From the Science for Life Laboratory,
Department of Biosciences and Nutrition (BioNut),
Karolinska Institutet, Stockholm, Sweden

DETERMINATION OF TRANSCRIPTION FACTOR BINDING SPECIFICITIES

Arttu Jolma

Stockholm 2015

Published by Karolinska Institutet.
Printed by E-Print AB 2015

ISBN 978-91-7676-122-9

Determination of transcription factor binding specificities
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Arttu Jolma

Friday 15th of December, 2015 15.00 in Novum, main lecture hall, Hälsövägen 7, Flemingsberg

Principal Supervisor:
Minna Taipale
Karolinska Institutet
Department of Biosciences and Nutrition

Co-supervisor:
Jussi Taipale
Karolinska Institutet and University of Helsinki
At the KI:Department of Biosciences and Nutrition; Science for Life Laboratory.
At the UH; Genome-Scale Biology Research Program; Institute of Biomedicine

Opponent:
John A. Stamatoyannopoulos
University of Washington, School of Medicine
Department of Genome Sciences

Examination Board:
Claes Wadelius
Uppsala Universitet
Department of Immunology, Genetics and Pathology
Division of Medical Genetics and Genomics

Rickard Sandberg
Karolinska Institutet
Department of Cell and Molecular Biology; Ludwig Institute for Cancer Research

Carsten Daub
Karolinska Institutet
Department of Biosciences and Nutrition

To all nice mammals

ABSTRACT

The term "genetic code" refers to the way in which the information encoded in nucleic acids is converted into the amino-acid sequence of proteins. There is however also a second genetic code, one that is used by the cells to read the blueprints of the entire organism. This second genetic code is composed of gene regulatory information that specifies how much of a gene product should be made when and where. This information is read primarily by sequence specific DNA binding proteins called transcription factors (TFs). TFs recognize and bind short DNA sequences that are located in the regions of DNA that are either just adjacent or relatively close to their target genes. When bound to these sites, TFs directly regulate transcription rates by recruiting the general transcription machinery, or by inhibiting its recruitment. Alternatively, TFs can influence transcription rates indirectly by recruiting proteins that will change the local chromatin environment in a way that will promote or inhibit transcription.

Each TF has its target specificity, it binds to a range of similar sequences that can be ranked based on their relative binding strengths. A major gap in our understanding of life is the lack of knowledge of the TF DNA binding-specificities. While we have good estimates of the total number of TFs and their general types, we do not yet understand the way in which the gene regulatory instructions are encoded in the genome. To approach this important question, we first need to know which DNA sequences TFs bind and how strongly.

The aim of this thesis project was to develop efficient methods for the characterization of TF binding specificities and then use these methods to catalogue DNA-binding specificities of as many human TFs as possible.

In Study I, we converted the classical Systematic Evolution of Ligands by Exponential Enrichment (SELEX) assay into a high throughput compatible method (HT-SELEX) and showcased the method by analyzing DNA binding specificities of 18 TFs representing 14 structural classes. Some of the results were validated by *in vivo* results from chromatin immunoprecipitation assays.

In Study II, we used HT-SELEX to analyze the binding specificities for clones representing almost all human TFs, generating a dataset of high resolution DNA binding specificity models for more mammalian TFs than in the entire previously published literature combined. Another major feature of our dataset is its high consistency, which was achieved by performing all of the experiments in parallel with the same method.

In Study III we studied evolution of gene regulation by analyzing the DNA binding specificities of TFs from the fruit fly *Drosophila melanogaster*. Analysis showed that even though the common ancestor of human and insects lived over 600 million years ago, the TF binding-specificities were very conserved between these species and there were similar counterparts to almost all of the TFs in either of the species.

LIST OF SCIENTIFIC PAPERS

Papers included in this thesis

I. **Jolma A**, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpää MJ, Bonke M, Palin K, Talukder S. Hughes TR, Luscombe NM, Ukkonen E, and Taipale J. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. 2010 Jun; 20(6): 861-73.

II. **Jolma A***, Yan J*, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T and Taipale J. DNA-binding specificities of human transcription factors. *Cell.* 2013 Jan 17; 152(1-2): 327-39.

III. Nitta KR, **Jolma A**, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, Taipale J. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*. 2015 Mar 17;4.

Additional papers

IV. Wei G, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Martin Bonke, **Jolma A**, Varjosalo M, Gehrke AR, Yan J, Talukder S, Turunen M, Taipale M, Stunnenberg HG, Ukkonen E, Hughes TR, Bulyk ML, Taipale J. Genome-wide analysis of ETS- family DNA-binding *in vitro* and *in vivo*. *EMBO J*. 2010 Jul 7; 29(13): 2147- 60.

V. **Jolma A** and Taipale J. Methods for Analysis of Transcription Factor DNA-Binding Specificity *in vitro*. *Subcell Biochem*. 2011;52:155-73.

VI. Whitington T, **Jolma A** and Taipale J. Beyond the balance of activator and repressor. *Sci Signal*. 2011 Jun 7;4(176).

VII. Yan J*, Enge M*, Whitington T, Dave K, Liu J, Sur I, Schmierer B, **Jolma A**, Kivioja T, Taipale M, Taipale J. Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell*. 2013 Aug 15; 154(4): 801-13.

VIII. Huang Q, Whitington T, Gao P, Lindberg JF, Yang Y, Sun J, Väisänen M, Szulkin R, Annala M, Yan J, Egevad LA, Zhang K, Lin R, **Jolma A**, Nykter M, Manninen A, Wiklund F, Vaarala MH, Visakorpi T, Xu J, Taipale J, Wei G. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nature genetics* 46 (2), 126-135.

IX. **Jolma A**, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E and Taipale J. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 2015 10.1038/nature15518.

X. Morgunova E, Yin Y, **Jolma A**, Dave K, Schmierer B, Popov A, Eremina N, Nilsson L and Taipale J. *Nature Communications* (in press).

* Authors contributed equally

Contents

List of abbreviations

3C	Chromatin conformation capture
B1H	Bacterial one-hybrid
bp	base pair
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CpG	DNA dinucleotide where C-base is followed directly by a G-base.
DBD	DNA Binding Domain (of a transcription factor)
DNA	Deoxy-Ribonucleic acid
FRAP	Fluorescence Recovery after Photobleaching
Hi-C	Massively parallel sequencing adapted 3C
HT-SELEX	High-throughput SELEX
IUPAC	International Union of Pure and Applied Chemistry
mCpG	methyl- CPG
mRNA	messenger RNA
NMR	Nuclear magnetic resonance
PBM	Protein binding microarray
PCR	Polymerase chain reaction
PWM	Position Weight Matrix
RNA	Ribonucleic acid
rRNA	ribosomal RNA
SBP	Streptavidin Binding Peptide
SELEX	Systematic evolution of ligands by exponential enrichment
TAD	Topologically Associated Domain
TF	Transcription Factor
TSS	Transcription start site
SELEX	Systematic evolution of ligands by exponential enrichment

Names of proteins, their structural classes and complexes

CBP	CEBPG-binding protein (a histone acetyltransferase)
CTCF	CCCTC-binding factor (incorrectly, CCCTC is not sufficient)
DNMT[1/3a/3b]	DNA methyltransferases 1, 3a and 3b
ETS	E twenty-six protein like family (homologous to a TF from a bird virus)
FOX	Forkhead (from appearance of a associated Drosophila phenotype)
FOXA1	Forkhead box A1
GATA	GATAA binding factor
H3K[4,27]	Histone modifications, e.g. H3K27 is histone 3, lysine 27
HMG	High mobility group
HP1	Heterochromatin protein 1
HSF	Heat shock factor
IRF	Interferon regulatory factor
NRE	Negative regulatory elements
NRF1	Nuclear respiratory factor 1
NURD	Nucleosome remodeling and deacetylation
p-TEFb	kinase protein complex associated with elongation
PHD	Plant homeodomain
PIC	Pre-initiation complex
RFX	Regulatory factor X
RNApolII	RNA polymerase II
SOX	Sry-related HMG box
SRY	sex-determining region Y, a HMG TF
SWI/SNF	Nucleosome remodeler
TAF	TATA- binding protein associated factors
TALE	Triple amino acid loop
TBP	TATA-binding protein

IUPAC code of degenerate DNA bases

R = A or G	B = G,C or T
K = G or T	D = A,G or T
M = A or C	H = A,C or T
W = A or T	V = A,C or G
S = G or C	N = A,C,G or T
Y = C or T	

1 Sequence specific transcription factors

1.1 INTRODUCTION TO GENE REGULATION

All life requires systems that regulate the amounts of their various protein- and RNA molecules, and they need to be able to respond to external stimuli by changing the amounts of these macromolecules. Products of the genes are either ribonucleic acid (RNA) molecules that are functional by themselves, such as ribosomal- or transfer RNAs, or they are messenger RNA (mRNAs) molecules that are subsequently transported out of the nucleus and translated into proteins.

Most of the directly functional RNA molecules are general cellular components and mainly due to this their regulation is a much more simple process than that of many protein coding mRNAs. They are produced with dedicated RNA-polymerase I and III based regulatory systems, of which the RNA pol I centered system produces the large ribosomal RNAs[1] and the RNA pol III based system makes most of the small RNA molecules such as transfer RNAs[2]. All of the protein coding genes and a subset of the functional RNA molecules are however made using RNA polymerase II, which is controlled through highly complex regulatory systems.

Cells use multiple mechanisms to control their protein levels; 1) Are the gene regulatory elements accessible for the transcriptional machinery? 2) How often those accessible genes are read into mRNA? 3) How much of the produced mRNA reaches the ribosome and is translated to proteins? 4) And finally, the proteins have different stabilities and their number is often controlled through selective breakdown[3].

The first two stages are controlled mostly by sequence specific DNA binding proteins called transcription factors (TFs). TFs function by recognizing and binding to target sites that are located in gene regulatory regions, relatively closely to their target genes. Gene regulatory regions contain typically tens of target sites for multiple different TFs, and the regulatory outcome is determined by the combination of TFs that bind to it. TFs that increase the expression of their target genes are called activators and those that decrease it are called repressors. Both of them can operate through accessibility or expression rate based mechanisms. Some TFs recruit chromatin modifying protein complexes that remodel the local chromatin towards a more open or closed state, and others control the level of gene expression by recruiting components of the general transcriptional machinery to transcription start sites (TSS)[3]

Chromatin accessibility (chromatin state) based control is especially important for multicellular life, as many of the genes are relevant only to certain cell types. Consequently, much of the transcriptional regulation related to cell differentiation occurs at this level[4,5]. Control of the transcription rate on the other hand is not associated as strongly with cell differentiation, but more to dynamic control of cells functions and responses to external stimuli (Reviewed in[6]).

Proteins need to be present in cells in suitable numbers, which varies enormously depending on the type of protein, and the cell type. While some proteins are expressed at very low numbers (tens to hundreds of copies per cell), others are expressed in scales of tens of millions[7]. In most of the cases the expression of wrong proteins would be just wasteful, but in other instances the misplaced expression of proteins, especially certain TFs, can lead to highly adverse outcomes such as reversal of cell differentiation or differentiation of the cell to a wrong lineage, which can then manifest in developmental defects or cancer. Thus, the gene regulatory system has to be both able to code for highly different levels of expression and to be very specific.

The genome contains approximately 20,500 true protein-coding genes[8], however their coding parts take only approximately a fraction of 1% of its entire length[9]. This does not mean that all of the rest is "junk". Up to a fraction of 10-20% is likely to contain gene regulatory information based on multiple lines of evidence such as evolutionary conservation, regions associated to diseases or phenotypic traits, as well as direct evidence based on high-throughput analyses[10,11].

Instructions about protein expression levels may sound uninteresting to a layperson, but as those regions also contain the blueprints of the entire animal, an understanding of the code or "language" of gene regulation would be the way to answer the ages old question; what separates mice from men?

1.2 REGULATION OF THE CHROMATIN STATE

The entire human genome is composed of over three billion (>3,300,000,000) DNA base pairs. It is almost two meters long chain of atoms and yet it is composed of only 23 massive molecules known as the chromosomes. Remarkably, two copies of the genome are packed into the nucleus, an ellipsoid with a diameter of only approximately 10μm. This tight packaging leads to an extremely crowded environment. It has been estimated that the total concentration of macromolecules, DNA, RNA and proteins, is in the range of over 100-200 mg/ml[12], which is higher than the solubility limit of many of the individual proteins. Furthermore, this packaging needs to be carried out in a efficient manner; the genome must not get entangled and all of its functional regions need to be situated in appropriate locations of the nucleus to allow precisely controlled transcription of all required genes. Unless the cell is one of the types that have lost their ability to divide, such as certain neurons, the genome packaging must also be capable of ordered disassembly in order to replicate the genome[13].

Unlike transcriptional regulation in simple organisms such as bacteria, where a single transcription factor target site can regulate multiple genes, the regulation of transcription in multicellular organisms is very complex. In eukaryotic organisms the length of gene regulatory elements is typically from hundreds to thousands of bases and these regions contain target sites for tens of transcription factors. Furthermore very little of the eukaryotic DNA occurs in naked state, but most of it is packaged around histone proteins to form nucleosomes, which is the smallest unit of the functional genetic material called chromatin[3].

The chromatin state, the extent to which the DNA is packaged by binding to nuclear proteins, is used as a gene regulatory strategy in eukaryotes, as the degree of packaging makes these regions harder or more easily accessible by both TFs and the basal transcriptional machinery. The chromatin state is controlled on many levels all the way from individual nucleosomes to up to the organization of the entire nucleus[14].

Chromatin can be roughly classified into two different categories, inactive heterochromatin that is by far the more common type in human, and active euchromatin that is limited mainly to genes and regulatory elements that are active in that cell. “Silenced chromatin” does not necessarily mean that the associated genes are also silenced, as the silenced region can just as well be a negative regulatory element, whose silencing activates the gene it normally represses[14,15].

Both hetero- and euchromatin are associated with nucleosomes that are spaced by 10-70 bases of free DNA, but the density of this packaging is different and both chromatin types bear their own distinctive epigenetic markings, which can be either covalent modifications of DNA by methylation of the cytosines located in the CpG dinucleotides, or through modification of the N-terminal regions, the “tails” of the histone proteins[14,15].In heterochromatin, the nucleosomes are separated by only very short and evenly spaced gaps of DNA and the tails of the different nucleosomes contact each other through protein-protein

interactions leading into further wrapping of the entire thread to form a 30 nm thick chromatin fiber in which only very little of the DNA is exposed to the external environment. In contrast, in euchromatin the spacings between the DNA bound histones are much longer and of more variable length, resembling beads on a string rather than the thick fiber of the heterochromatin[16].

Some genomic regions have the same chromatin state in essentially all types of cells. For example, genes that serve general housekeeping duties are in a euchromatic state in practically all cells. Regions that are in a heterochromatic state in all cell types are called constitutive heterochromatin, this is in contrast to facultative heterochromatin, whose state is different depending of the cell type. Constitutive heterochromatin includes structural regions such as telomeres, the "protective caps" located at the end of the chromosomes; centromeres, which are used to move and bind together chromosomes during mitosis as well as repetitive regions that are often derived from parasitic genomic elements. Facultatively heterochromatic regions on the other hand are packaged either into hetero- or euchromatin depending on the cell type and include regulatory and coding parts of genes. In general, the less differentiated cells have more euchromatin, which during differentiation gets locked up more and more into a heterochromatin state[17].

1.2.1 DNA CpG-methylation

Around 90% of all CpG (C followed directly by G) dinucleotides are methylated in human cells. Besides CpG dinucleotides, methylated cytosines can occur also in different sequence contexts. Non-CpG methylation (most commonly CAG) exists in mammalian pluripotent stem cells of both embryonic and artificially induced varieties, but is lost when the cells differentiate. This is likely because the non-CpG methylation is not actively maintained with a specific enzyme like CpG methylation (see below)[18]. The methylated CpG dinucleotides are not distributed evenly in the genome; in general, heterochromatin is more methylated than euchromatin, and the degree of methylation also varies within these regions.

CpG is methylated by DNA methyltransferase enzymes (DNMT:s) that either methylate the cytosines of the CpG residues *de novo* (DNMT3A, and DNMT3B), or function as a maintenance-methylase (DNMT1). The latter maintains CpG methylations by recognizing the positions where only one of the two cytosines on the sister strands is methylated, a situation that commonly occurs after DNA replication. The effect of CpG methylation depends on the context: methylation of regulatory DNA-regions has usually a silencing effect. Heavily methylated regions are packaged tightly into heterochromatin, where the internucleosomal regions are even more highly methylated than the parts that are wrapped around the core histones[19]. On the other hand, mCpGs occur also commonly in the gene bodies of active genes, which is thought either to serve as a way to suppress spurious initiation of transcription from within the genes, or to be simply a passive feature of the relatively open chromatin structure that makes it easier for the DNMT-enzymes to access their substrate[20].

CpG methylation can be reversed passively as the replication cycles of the genome will initially generate unmethylated DNA, until methylation of the daughter strands by DNMT1. Active mechanisms on the other hand are based on a multi-stage process: in the first step, so called TET-enzymes catalyze the oxidation of the methyl group, first to hydroxymethyl, which can then be oxidized further to produce 5-formylcytosine and finally 5-carboxylcytosine. In the second step, the entire oxidized methylcytosine base is removed by base excision repair. There is also some evidence for other cellular processes that could be used to reverse the methylations and restore the DNA base back to its original state, but so far their existence is controversial[21].

CpG methylation is used as a general mechanism in the regulation of the chromatin state and as a way to modify the transcription rates of genes, but it is also used in a few special cases such as in the inactivation of the entire second X-chromosome in female mammals, as well as in silencing of parasitic DNA such as endogenous retroviral elements[22]. Furthermore, there is at least a single relevant context, imprinting, where the mCpG based epigenetic modification states are even passed from one generation to the next. In imprinting, certain genomic regions are silenced if they were inherited either from the mother or the father. Imprinting has been estimated to affect expression of 200-1300 mouse genes (around 1-6.5%)[23].

The mCpG based silencing of gene regulatory elements is thought to be mediated primarily by proteins that have dedicated methyl-CpG binding domains (MBDs). MBDs bind to methylated CpGs regardless of the sequence context of the flanking regions and recruit proteins that modify histone tails or remodel the associated regions into heterochromatin [19]. CpG methylation has also a direct effect on the binding of many sequence specific TFs. Many TFs can bind sequences that contain a CpG dinucleotide, and the methylation of these sites can affect the binding affinity of the TF, in most of the known cases this effect is negative, for example binding of certain TFs from the AP2[24]-, bZIP[25]- and bHLH[26] structural families has been shown to be inhibited by CpG methylation of their target sites. In some cases, TFs can however bind mCpG methylated sites better than the unmethylated variant or even recognize novel mCpG-sites that would not be bound in the absence of methylation[27].

1.2.2 Histone modification, assembly and remodeling

Histones are multiprotein complexes composed of four to five types of protein subunits. The central ”core” part is functionally like a spool made of positively charged protein molecules that binds to the negatively charged DNA backbone leading to coiling of 147 bp of DNA around it. Core histones are composed of eight protein molecules, two copies each of the H2A, H2B, H3 and H4 subunits. Besides this octameric core, many nucleosomes, especially the ones located in heterochromatin, have an additional protein part known as the linker histone H1, located at the points where the DNA exits the core region. Cells express many variants of the five different types of histones, some of which have been shown to have important roles in gene regulation, e.g. the variant form of histone H2A, H2A.Z which occurs predominantly in the nucleosomes that flank the open regions of the genome and is thought to decrease the strength of the nucleosome-DNA interaction[28].

Both DNA replication and transcription remove all histones from the DNA, after which they are loaded back into the DNA by histone chaperones[15,29]. Loading of the histones is likely to be carried out in two phases, where a tetramer composed of two H3 and two H4 subunits is loaded first, followed by loading of H2A-H2B heterodimers, and this happens without external energy sources[30].

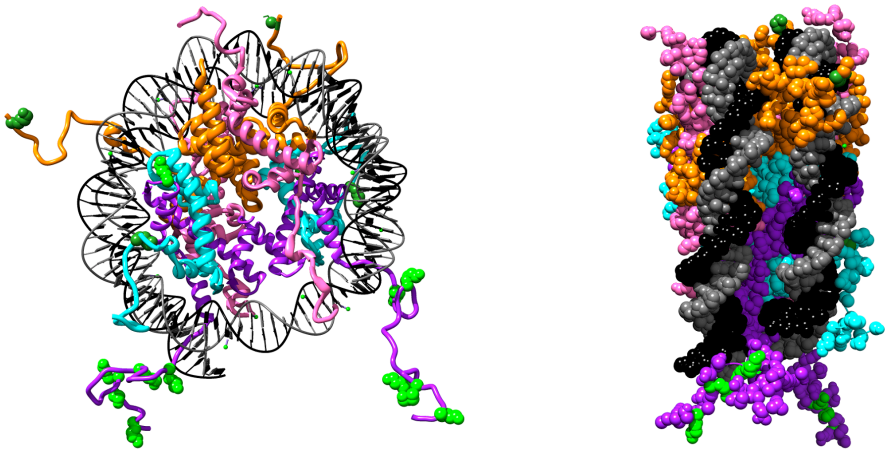


Figure 1 | Nucleosome
Figure shows the crystal structure of a nucleosome displayed as two different representations, as a cartoon (left) and a space filling model (right), the latter of which has been also turned by 90 degrees counter-clockwise. Four histone subunit proteins are marked with different colours (H2A, pink; H2B, orange; H3, violet; H4, cyan) and the commonly modified lysine residues have been marked as space-filling green representations in both of the views. Note the multiple modifiable positions on the tails of violet H3 and that they are located in the positions where the DNA enters and exits the nucleosome spool. The structure has been visualized using UCSC Chimera[31] from PDB structure 1KX5[32].

1.2.2.1 Histone modification

Multiple amino acid residues located on the N-terminal "tails" of the four core histone proteins can be modified through covalent addition of side groups. Different amino acid residues of the H3 and H4 proteins can be at least acetylated, methylated, phosphorylated, sumoylated or crotonylated[14,29]. In addition, the tail region of H3 can be cleaved off by a specific protease[29] H2A and H2B proteins on the other hand are limited to ubiquitination of a single position. Some of the histone modifications, such as phosphorylation and acetylation can directly affect the interaction between the histone and the DNA, as they change the charge of the affected amino acid residues[14]. Different histone modifications can change the accessibility of the underlying DNA. Based on nuclease protection assays, the modifications can either decrease or extend the canonical nucleosome-contacted length of DNA from ~147 bases to 106–164 nuclease protected bases[33]. A detailed analysis of all possible modifications is beyond the scope of this work and this introduction focuses instead on a few prominent examples.

Acetylation of lysine residues can occur on multiple positions of the H3 and H4 proteins and all these modifications activate chromatin. Most of the enzymes catalyzing acetylation, the histone acetyl transferases (HATs), are promiscuous and can acetylate many different lysines of the histone tails. Addition of an acetyl group to a lysine residue neutralizes its positive charge and can thus have a relatively strong direct effect on the contacts between histone and the DNA backbone. Besides the direct effect to the histone-DNA interaction, these modifications function also as a way to recruit chromatin-modifying enzymes. The acetylated lysines can be recognized by protein folds known as the bromodomain and the plant homeodomain (PHD), which occur for example in several histone-remodeling enzymes[15]. Conversely, histone acetylation can be reversed by histone deacetylases (HDACs), which is usually connected to silencing of the genomic region. The connection of histone deacetylation to silencing is however not total. While promoters of active genes are often acetylated, the coding regions contain both acetylated and deacetylated histones, and it has been suggested that this is connected to elongation of the transcript in a dynamic fashion serving the same function as the CpG methylation of the gene bodies. Based on this theory the histones would be acetylated in front of the RNA polymerase to help the transcriptional elongation, after which they would be deacetylated again to inhibit sporadic initiation of the transcription from within the coding regions of the genes[34].

Histone methylation can occur on either lysine or arginine residues of H3 and H4, and this modification can have either activating or silencing effect depending on the amino acid residue that is modified within the histone tail. The most common activating methylations are mono-, di- and tri- methylations of lysine 4 of the histone 3 protein (H3K4Me[1,2,3]) and the most common silencing ones are the mono-, di- and tri- methylations of lysine 27 of histone 3 (H3K27Me[1,2,3]). The distinct effects of different methylations are also mirrored by the enzymes that add or remove these groups, as in contrast to the promiscuous HATs and HDACs, the histone- methyltransferases and demethylases are position-specific. Histone methylations do not change the charge of the affected residues and thus have no direct effect on the interactions between the lysines and the DNA backbone. Their effect is instead dependent on proteins that recognize and bind to these modified positions. Different methylations can cause very specific effects, for example the trimethylation of lysine 36 of histone 3 (H3K36Me3) is used to specifically guide and promote DNA CpG methylation of nearby bases through the action of *de novo* methylases DNMT3[A,B]. At least ten different kinds of protein domains have been shown to be able to recognize specific methylated positions on the histones[14].

1.2.2.2 Histone remodeling

Besides the passive histone loading processes facilitated by the histone chaperones, the core histones can be loaded, moved around or ejected actively from the DNA by many multiprotein complexes called chromatin remodelers that modify chromatin structure using ATP as their energy source. There are at least ten distinct remodeling-complexes in human, and the actual diversity is even higher because of multiple paralogues for many individual

subunits. Some of the chromatin remodelers are used mainly in activating processes, while others are specialized in silencing and they are recruited to their target regions either by recognition of modified histones or by direct action of sequence specific TFs. Nucleosomes can be repositioned by sliding them along the DNA-strand, a process which can both package the histones into tight and evenly spaced arrays or reverse the tight packaging. Some chromatin remodelers can also catalyze the replacement of constituents of the nucleosomes, reducing the octameric nucleosome to tetrameric or even ejecting the nucleosomes altogether from the DNA[15,35]. Histone modification, especially histone deacetylation, and chromatin remodeling are tightly connected to each other. For example, one of the remodeling complexes, the nucleosome remodeling and deacetylation complex (NuRD) contains an mCpG recognizing subunit that is connected with domains with histone deacetylase and histone remodeling activities and thus both of these silencing functions are performed simultaneously in an mCpG guided fashion[34,35].

1.2.2.3 Nucleosome positioning

Conceptual separation of chromatin into only two kinds of states is a very rough generalization. Although heterochromatin is indeed often packaged into a very tight, almost crystalline state, euchromatin is more like a wide and heterogeneous continuum of more or less accessible states. Recent high throughput studies using mainly DNAseI hypersensitivity assays and ChIP-seq with anti-histone antibodies, have shown that the transcription rate is commonly regulated through dynamic competition between TFs and nucleosomes. It has been shown that the transcription cannot be initiated if a nucleosome is binding too closely to the TSS. Moreover, a suitably placed nucleosome at a distal regulatory element can inhibit contact formation with its target promoter[36]. Promoter core regions, the areas immediately adjacent to the TSS, are cleared of nucleosomes in active genes where the average nucleosome free area is 200bp long, with 80 and 120 bp distances between the TSS and the first up- and downstream flanking histones, respectively[28,33]. Positioning of the nucleosomes is affected by many different things, such as direct competition against DNA bound TFs, the presence of RNA polymerase II, CpG methylation, chromatin modifying/remodeling complexes and even by the inherent sequence specificity of the nucleosomes themselves (See below)[36].

Nucleosomes are not sequence specific in the same sense as TFs; they do not define continuous DNA sequences that could be described well using a motif representation. Instead their specificity is more about the local shape and deformability of the DNA and their specificity manifests as preference for certain dinucleotides in an approximately 10bp periodic pattern, as well as strong disfavoring of certain DNA sequences such as stretches of A or T nucleotides. Even though none of the individual DNA binding regions of the nucleosome footprint has a strong determining effect on its positioning, the sum of all individually weak positions over the entire 147 bases wide footprint of the nucleosome can yield strong net influence[37].

1.2.2.4 Inducers of active (open) chromatin state

Chromatin state activating TFs can function through many different mechanisms. Even though binding of essentially any TF will by itself compete against the binding of the histones, it is clear that in most cases the major effect is mediated by enzymes that are recruited by the TFs activation domains. These enzymes include mCpG demethylating TET-enzymes[21], histone modifiers such as histone acetyltransferases like p300/CBP, histone methyl-transferases (activating types, e.g. H3K4 specific methyl-transferases) or other chromatin remodelers that modify the chromatin towards the more open state[5,35].

Some of the activating TFs seem to be especially suited for binding to tightly packed heterochromatin, and these "pioneer factors" have been suggested to kickstart the activation of silent chromatin. Moreover, some pioneer factors have been suggested to not even require cofactors but to be sufficient to open compacted chromatin as their intrinsic property[38]. TFs belonging to the forkhead (FOX) structural class are for example thought to be particularly suited for these tasks: Their DBD is structurally highly similar to H1 linker histones[39], and can, like H1, access the major groove of histone bound DNA that is facing the nucleoplasm rather than the core histones. Unlike the linker histones, FOX TFs however recognize specific sequences and if such a site is present at the nucleoplasm facing major groove of the DNA, the TF will bind to it, causing eviction of the histone. Some of the FOX TFs like FOXA1 also contains a specialized C-terminal domain that can bind to the core histone proteins of the nucleosomes[40]. Other TFs suggested to function as pioneer factors include other forkhead TFs, various GATA TFs[38], and POU5F1, one of the Yamanaka factors. POU5F1 appears to be a particularly potent pioneer factor, as it functions in rebooting the cell all the way to pluripotent stem cells[41].

1.2.2.5 Inducers of silenced chromatin state

TFs that silence gene regulatory regions recruit corepressor complexes that have an essentially antagonistic set of enzymatic affinities when compared to activator TFs. These include HDACs, certain histone methyltransferases (inactivating, e.g. H3K27 specific methylases) and different sets of chromatin remodeling ATPases.

In early development, many lineage defining genes are silenced by the action of polycomb repressor complexes (PRC) that are recruited to their promoters by stemcell specific TFs, such as POU5F1 (which can function both as an activator and a repressor, depending on the DNA sequence and the presence of other TFs), SOX2 and NANOG. The first of the recruited PRC-complexes is PRC2, which mediates its silencing effect through H3K27 specific methyltransferase activity. In further steps the methylated H3K27 recruits the PRC1 complex, which mediates tighter packaging of the local chromatin[42]. During differentiation it is common (and in many cases this is PRC2 mediated) that the gene regulatory elements are neither fully silenced nor activated, but remain in a "poised" state, where chromatin is accessible and an inactive form of RNA polymerase II is loaded onto the promoter, but it is

kept inactive through H3K27Me3 and associated changes. Once the cell differentiates, a subset of these promoters get fully activated to guide cell differentiation[34,42].

Some of the TFs have specialized into an "all or nothing" type of silencing of genomic regions, particularly zinc finger proteins that contain SCAN and/or KRAB effector domains. A very large fraction of the analyzed TFs of this type have been shown to bind to different kinds of parasitic genomic sequences such as endogenous retroviral elements[43], where they serve as a form of molecular immunity that is used during early development to silence these elements. The KRAB and SCAN effector domains recruit KAP1-protein, which in turn mediates silencing of the nearby genomic regions through further recruitment of chromatin modifying enzyme complexes, nucleosome remodeling and deacetylase (NuRD), SETDB1 (histone methyltransferase) and heterochromatin protein 1 (HP1)[44].

1.2.3 Composition and types of gene regulatory elements

The genomes of complex multicellular animals contain massive numbers of regulatory elements, in the scale of hundreds of thousands, making up to 10-20% of the total base content of the genome[10]. Based on systematic DNAse-seq analysis of 125 human cell and tissue types there could be close to two million gene regulatory elements[45].

Regulatory elements can be classified based on their location relative to target genes or by their typical roles. Transcription is started from the transcription start site (TSS), which is located from tens up to a couple of thousand bases (human average distance is 210) upstream of the coding sequence of the gene[46]. Regulatory regions around the TSS are called promoters. The promoters have typically two distinct regions; the promoter proximal region that is located from tens to a few hundred bases upstream of the TSS and contains binding sites for sequence specific TFs, and the promoter core-region that is located directly over the TSS and often contains some of the general promoter sequence elements, such as the initiator element (Inr) or the TATA-box that function as binding sites for general transcription factors (proteins that are used as components in either all or at least a large subset of transcription initiation events). A single gene has often multiple promoters, and the choice of which of them is used to regulate transcription varies between cell types[47]. Based on occurrence of DNAse I hypersensitive areas near TSS positions it has been estimated that there are between 75,000 to 150,000 promoters that drive expression of the human genes[45].

In addition to promoters, there are several types of gene regulatory elements that are positioned further away from the TSS (distal elements) such as enhancers, silencers, and insulators. Distal elements are located typically around tens of thousands, but up to hundreds of thousands of bases either up- or downstream of the TSS of their target gene(s). Distal elements function by forming contacts with either promoters (most enhancers and a subset of silencers) or each other (insulators). These contacts are used to guide the folding and packing of chromatin, and in the case of promoter contacts, to also modulate the expression rates of the associated genes by guiding the assembly of the basal transcriptional machinery, a massive multi-protein complex called Pre-Initiation Complex (PIC)[6]). The nature of the

regulatory element is often context dependent, and the same element can function either as a silencer, enhancer or even insulator depending on the TFs that are bound to it. Even the distinction between promoter proximal regions and distal elements is a question of threshold of distance between promoter and the TF target sites rather than a genuine difference. Moreover, proximal promoters of a pair of genes have been shown to effectively function as each other's enhancers[48].

Many of the regulatory elements do not fit the traditional classifications very well. A form of negative regulatory element can function effectively as a decoy to which an enhancer will bind instead of the target promoter. In blood cell development for instance, the expression of the gene KIT is suppressed by the TF GATA1, which guides its enhancer to contact a gene regulatory element downstream of the TSS instead of the KIT promoter[49]. Other kinds of indirect mechanisms could be for example based on two regulatory elements that would help to compress the span of chromatin between the actual enhancer-promoter pair, or on making a negatively functioning element by forming a bulky, sterically blocking loop adjacent to a promoter or enhancer[50].

Thus it is important to keep in mind that the classifications of the different kinds of regulatory elements are just rough generalizations made to facilitate their description using human language while their real character is often dependent on the cellular context.

1.2.3.1 Enhancers

An enhancer and the promoter that it modulates are connected to each other by looping and then folding and compacting out the intermediate region of the chromatin, and these two elements are usually located relatively close to each other. An bioinformatics study based on analyses of previously generated data from Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) assay performed with RNA polymerase II antibody, estimated that the median distance of enhancer-promoter pairs was 15,000 bases in the two analyzed cell lines[51,52].

An enhancer often controls the promoter closest to it, but in many cases the connecting stretch of chromatin can span hundreds of thousands of bases and contain multiple other genes. In a few extreme instances enhancers have even been shown to drive expression from promoters that are located on different chromosomes. For example, each smell sensing neuron expresses only one out of hundreds of possible olfactory receptor genes encoded in its genome, and these receptor genes are scattered into tens of loci in almost all chromosomes[53]. Choice of the olfactory receptor that the cell will express appears to be a stochastic process in which the enhancer gets connected to a promoter of a single olfactory receptor gene regardless of its location in the genome. Nevertheless, the likelihood of getting connected to the enhancer is higher for those receptor genes that are located in close positions of the same chromosome[54].

Traditionally the active promoter of a gene was thought to be connected to only one enhancer at a time, based on recent high-throughput studies it however seems that exceptions to this

rule are common; multiple enhancers can be connected to the same promoter in as often as 25% of cases. Many promoters can also be connected to the same enhancer simultaneously[48,55].

1.2.3.2 Silencers

Classification of the silencers is often ambiguous, as the term is used to describe many kinds of different elements such as negative regulatory elements (NRE) and traditional silencers. NREs are individual target sites for repressive TFs that are located in the promoters and enhancers and function either by blocking a DNA binding site of an activator, or by otherwise hampering the assembly of the PIC components by e.g. blocking a necessary protein-protein interaction interface of a general TF. Traditional silencer elements on the other hand are dedicated regulatory elements that have multiple TF binding sites and function similarly to enhancers by forming a contact with their target promoter, which however does not lead to transcription, as the silencer lacks several of the key binding sites required for transcriptional activation. Other kinds of silencer regions and elements are used to bind repressor TFs that operate indirectly through further recruitment of co-effectors, HDACs or DNA methyltransferases leading to modification of the local chromatin towards a suppressed state. These kinds of silencers are used to repress entire regions of chromatin during cell differentiation[5,6].

1.2.3.3 Insulators

Early findings showed that the effect of enhancers can be limited with gene regulatory elements known as insulators. These elements appear to inhibit the contact between enhancers and promoters, as shown experimentally by introduction of a new insulator element between the promoter and its enhancer[56]. From early on one of the main theories to explain the function of the insulators was that they are used to separate the genome into local domains. This has been now confirmed through chromatin conformation capture (3C) based high throughput experiments (Hi-C), which showed that the genome is organized into about megabase long locally folded regions called topologically associated domains (TADs). The TADs are visible in the Hi-C experiments as regions within which sequences contact each other much more commonly than they do in the outside lying regions, and thus are likely to represent chromatin that is packed into the same local structure. The beginning and the end of these regions are associated to each other even more commonly than the regions within, suggesting that the TAD is a loop of chromatin that is then folded further into a tight globule. Most of the enhancer–promoter interactions are limited to occur within TADs and their borders correlate well with both known insulator elements such as CTCF as well as the boundaries of the genome's division into hetero- or euchromatin[57,58]. Further upgrades to chromatin conformation capture methods have shown that the megabase TADs can be subdivided into even smaller locally folded regions, many of which have enhancer-promoter pairs located at the points where the ends of the domain contact each other. When taking together both the megabase level of organization (insulator dependent) and the local level folding of the genome that is connected to the transcriptional regulation, the human genome

seems to be organized into median locally folded compartments of a median size of 185 kilobases[59].

Unlike the enhancers and promoters that are composed of highly diverse sets of different TFs, the insulator elements contain a common sequence specific element that is recognized by the C2H2 zinc finger TF CTCF. This TF binds long and highly specific NGCGCCMYCTAGYGGTN target consensus sites that are, despite the length of the site, very common in the genome[60]. The prominent role of CTCF in genome organization is also highlighted by the fact that, similarly to the situation at the TSS, it can also position nucleosomes at defined spacings around its target sites[33]. The details of the mechanisms that tie the ends of TADs together are still unclear, but besides CTCF these contacts employ cohesin, which is a ring shaped multiprotein complex that can connect two nearby strands of chromatin by encircling them[59]. CpG methylation can be used as a way to mask the function of the insulator element by blocking the binding site of CTCF[61].

Figure 2 | CTCF binding specificity (From Study II)

1.3 THE GENERAL TRANSCRIPTIONAL MACHINERY

Transcription is initiated by the assembly of the Pre-Initiation Complex (PIC) that is composed of many general transcription factors (26 peptides on six complexes; TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, making altogether 1560 kD) and the RNA polymerase II (12 peptides, 515 kD). Experiments using isolated proteins have shown that these components of the PIC are sufficient to drive low level expression even in the absence of the mediator complex (composed of 21 peptides with a total molecular weight of ~1005 kD), but mediator is required for more efficient transcription, where it channels the effects of the activating TFs bound on enhancer or proximal promoter. Furthermore, mediator is also required for the efficient transition from the PIC phase to the following steps, the pre-elongation and then the actual elongation phase, where RNA polymerase II is released from its contacts with the rest of the PIC components in a controlled and stage-wise fashion. Many structures of the subunits of both PIC and mediator have been solved using x-ray crystallography and nuclear magnetic resonance (NMR), and recent research has even been able to generate low resolution cryo-electron microscopy based structures of partial PIC complexes, (see Figure 3)[62,63].

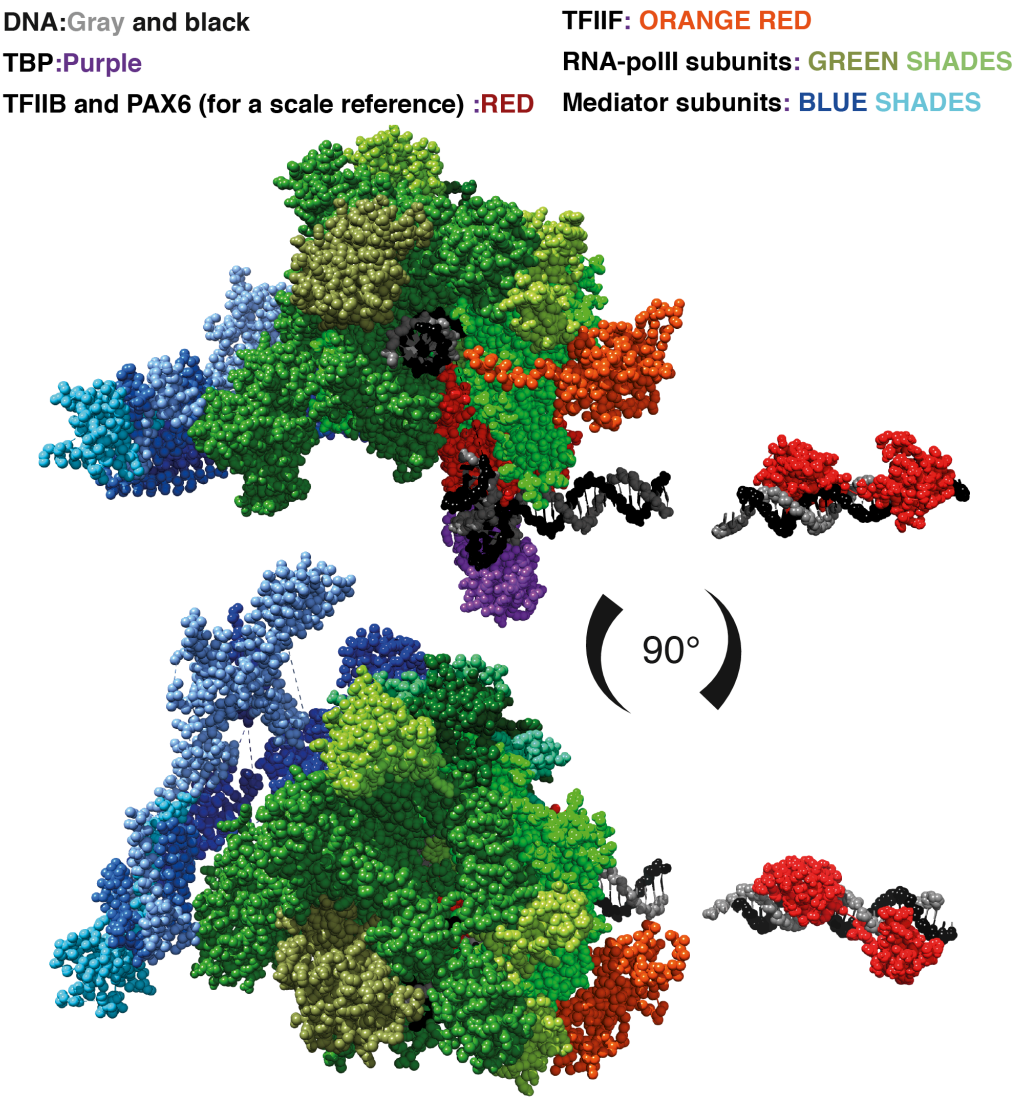


Figure 3 | Partial structure of mediator and RNA-Polymerase II
The figure shows a partial structure of the RNA-Polymerase II pre-initiation complex based on individual high resolution X-ray diffraction structures that were fitted to a low resolution cryo-electron microscopy based structure of the entire complex. The structure is shown from two angles, side and top, and has been visualized using UCSC Chimera[31] using the data from PDB structure 4V1O[63].

The function of the mediator is to tether relevant enhancer element(s) to the promoter and through this to mediate the effects of the TFs, such as recruitment of the general transcription factors or chromatin modifying complexes to the PIC. TFs that are bound to the paired enhancer and promoter recognize different parts of the mediator and the composition of the mediator is varied to some extent depending on the cell type[62,64]. Based on various sources of new data, including our laboratory's contributions, this tether is very commonly reinforced through action of cohesin, a very long oval-shaped protein complex that forms loop around the two strands of chromatin[55,65,66].

As initial stage of the PIC assembly, the TFIID, which is composed of TATA binding protein (TBP) and 10-12 other proteins called TAFs (TATA-associated factors), is recruited to the

promoter. Unlike most of the other elements of the PIC or mediator, the TFIID has inherent DNA and chromatin recognizing activity. It can bind to the TATA-element (consensus TATAAWR) with the TBP[47], and a complex formed of two of its subunits TAFI and TAF2, which together recognize the second commonly occurring core promoter sequence, the initiator element (Inr, with YYANWYY consensus)[67]. TFIID components can also read epigenetic information, as the subunit TAF1 has two bromodomains that recognize acetylated positions on histone H4[68]. Similarly, TAF3 has a plant homeodomain (PHD) that it uses to bind to H3K4me3[69].

Next, is TFIIB binds upstream of TFIID, making contacts with both TFIID and the DNA. This assembly stage of the PIC is sufficient to bind RNA polymerase II and defines the start site for the transcription, but initiation will not occur until the remaining complexes TFIIE and TFIIH bind to the PIC (reviewed in[62,70]).

It is unclear to which extent the PIC gets disassembled once elongation has begun and this probably varies on different promoters. It has been suggested that almost everything except for the TFIID is disassembled, or on the other extreme that even the mediator and the TFIID, TFIIE and TFIIH would be retained on their promoter associated positions to form a scaffold for the rapid reassembly of the PIC[6]. PIC is a very large multiprotein complex that is connected to DNA through many of its subunits and due to this its passive disassembly would be too slow for the regulatory needs of some promoters. One of the mechanisms used to facilitate the disassembly is through action of the ubiquitin system, which marks up some of the PIC components for proteasomal degradation. In other cases, the components can also be removed through action of nucleosome remodelers such as SWI/SNF complexes[5].

At many promoters, cells employ yet another layer of transcriptional control where transcription is paused after synthesis of the first 30 to 50 bases. Both the pausing and its reversal are controlled through interactions between the RNA polymerase II and additional proteins called elongation factors. Transcriptional pausing can be caused by the barrier presented by the first nucleosome, the phosphorylation stage of the C-terminal part of the active subunit of RNA polymerase II, or by tethering the RNA polymerase II to the promoter region by the action of DNA binding proteins. Sequence specific TFs are also commonly used in the control of theses stages. The MYC-MAX heterodimeric TF complex for instance promotes transcriptional elongation of many of its target promoters through recruitment of a kinase protein complex called p-TEFb that can phosphorylate the C-terminus of the RNA polymerase II and many other associated proteins[71].

1.4 GENERAL CHARACTERISTICS OF TFs

Most TFs recognize quite small, on average about 11 bases long target sites. Each specific sequence of that length occurs approximately 700 times in the human genome. Furthermore, many of the positions within those target sites are not strictly defined, but have an average information content of 1.2 bits per base[72] making a total of 13.2 bits. As the information content of an absolutely defined base is 2 bits, this means that an average TF defines only 6.6

full bases worth of information and thus the practical number of the binding sites is much larger than the 700 exact matches, in the order of tens to hundreds of thousands of sites[72,73]. TFs are essentially the DNA recognizing interface of gene regulation and thus their main function, besides DNA binding, is to bind to other proteins to recruit them to the regulatory regions of the genes.

1.4.1.1 Regulation of TF activity

TFs themselves are typically the targets of cellular signaling cascades and are regulated by different kind of mechanisms that are based for example on: 1) Expressing the relevant TFs when they are needed. 2) Activating or silencing those TFs that are already present in the cell. 3) Controlling the amount of the TFs in the cells through targeted degradation.

TFs can be in an inactive state due to many different reasons, for example the ATF6 and CREB3 type bZIP TFs are normally bound to the membrane of the endoplasmic reticulum and this membrane-binding domain has to be cleaved off to activate them[74]. As another example, all STAT TFs need to be activated by phosphorylation in order to translocate into the nucleus, multimerize and then bind DNA[75] and most nuclear receptors need to bind to their target substrates in order to be able to bind DNA[6]. Many of those nuclear receptors, such as estrogen-, Vitamin-D- and progesterone- receptors are also controlled through active ubiquitin guided degradation systems to get rid of ligand bound TFs so that the cells can assay the concentrations of their target ligands accurately[5].

1.4.1.2 Effective copy numbers of TFs in cells

TFs are expressed in relatively low numbers when compared to the general expression levels of proteins, for example a mass spectrometry based analysis of the human osteosarcoma cell line U2OS detected 36% of the estimated total of 20,000 proteins, and 23% were present in more than a 1000 copies. The total number of detected TFs was however only 20% of all approximately 1300 TFs, and only 7% were present in more than 1000 copies[7]. Recent genomics analyses suggest that the identity of a cell type is typically determined by expression of only a few master regulator TFs that are expressed in very high copy numbers, 250,000-500,000 copies, which is in contrast to the rest of the around 200 TFs that are expressed in twenty- to fifty fold lesser numbers ~10,000 copies per cell[7].

1.4.1.3 TF effector domains

In the simplest case, a TFs consists only of their DNA binding domain (DBD), such a minimal TF can either repress expression of an activator TF by simply competing for binding into the same site, or it can exert its function by cooperating with another TF through protein-protein or DNA shape -mediated interactions. An example of such a minimal TF is BATF that is composed of only 125 amino acids, and functions through forming heterodimers with other bZIP TFs[76].

Most of the TFs however contain dedicated effector domains besides their DBDs, which are usually connected to the DBDs by flexible linker regions. Effector domains are often simple

protein–protein interaction interfaces that are used to recruit several kinds of other proteins such as: 1) Co-activators and co-repressors, which are typically fairly large protein complexes with chromatin- modifying and/or remodeling activities. 2) Components of the basal transcriptional machinery, such as general TFs of the PIC or mediator components[5,77].

Many of the extensively researched TFs have fairly promiscuous effector domains that can form contacts with multiple different interaction partners, for example the activation domain of E2F1 has been shown to make contacts with many general TFs of the PIC[5,77].

Effector domains of most TFs have been characterized only very superficially. Some effector domains contain many acidic residues (aspartate and glutamate. e.g. E2F1 and TP53), while others are glutamine- (POU2F1, POU2F2 and SP1) or proline-rich (TFAP2A and NFIC). The linker region that connects the effector domains to the DBD is usually very flexible and it is likely that such unstructured linker regions are required because the TFs and the various proteins recruited by them will need to be able to make contacts with each other on the three dimensional environment around the sterically blocking DNA. Based on NMR experiments, most of the analyzed effector domains appear to be relatively unstructured when occurring free in the solution phase and only gain folded shape when binding to their target proteins[77].

Some of the TF effector domains are ambivalent; they can function either as an activator or repressor depending of the context. In many cases a TFs role is dependent on additional factors such as availability of cofactor(s) and the presence of other bound to nearby sites. In some cases it depends on the site that the TF is bound to, functioning as repressor on some and as an activator on others. This is thought to be mediated through allosteric effects, where binding of the TF to different sequence motifs leads to divergent changes on their fold, which then affects the properties of their protein–protein interaction interface. Examples of this kind of TFs include POU1F1, NFKB1 and several TFs from the nuclear receptor family[5,78].

1.5 TF STRUCTURAL FAMILIES

Evolution has developed sequence specific DNA binding domains in many independent instances and some of these basic protein folds diverged to subfamilies so long ago that their relatedness to one another is not reflected anymore in their DNA binding specificities. For example, the ETS, FOX, RFX and even the linker histones of H1 type are all based on the winged helix-turn helix protein fold, but since each of these proteins binds sites that are very distinct from each other, or, in the case of the linker histones, have no specificity at all, it is not reasonable to discuss them here in the same context. Thus even though the classification scheme used here is based on evolutionary relationships of the TFs, it uses the DNA recognition specificity rather than e.g. a strict amino-acid similarity threshold as the basis for their classification into families[72,79].

Interestingly, some of the TF families have expanded to large numbers of members while at least in one case, the nuclear respiratory factor, the structural family has only a single TF in human (NRF1)[79,80]

To complicate things further, some of the TFs, such as several members of the Homeodomain- PAX-, POU- and CUT-families, contain two different kinds of DBDs that are often used in different orientation and spacing combinations[72,73].

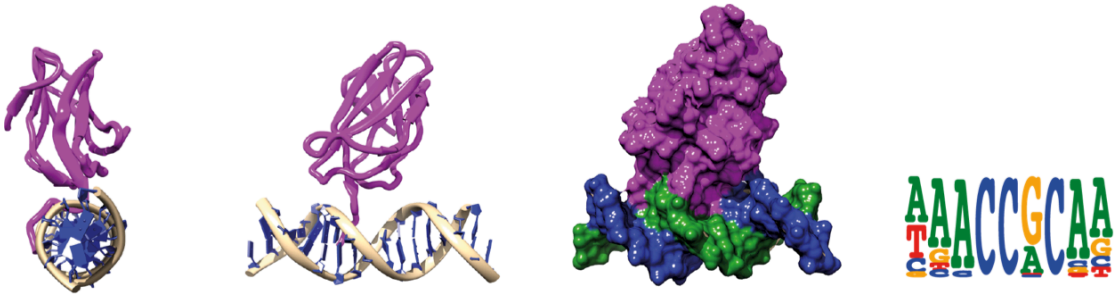
Name	Nr. of TFs	Interpro Id.	Name	Nr. of TFs	Interpro Id.
znfC2H2	745	IPR015880	SMAD	8	IPR013019
Homeodom.	243	IPR001356	CUT	7	IPR003350
bHLH	115	IPR011598	RFX	7	IPR003150
bZIP	56	IPR004827	STAT	7	IPR013801
HMG	54	IPR009071	DM	7	IPR001275
FOX	49	IPR001766	CP2	6	IPR007604
Nuc.Rec.	46	IPR001628	MADS	5	IPR002100
ETS	27	IPR000418	HSF	5	IPR000232
T-box	17	IPR001699	AP2	5	IPR013854
POU	16	IPR000327	TEAD	4	IPR000818
Gata	15	IPR000679	NFI	4	IPR020604
E2F	11	IPR003316	RUNX	3	IPR013524
RHD	10	IPR011539	p53	3	IPR011615
PAX	9	IPR001523,	GCM	2	IPR003902
IRF	9	IPR001346	NRF	1	IPR019526
SAND	9	IPR000770	Total	1505	

Table 1. Structural classes of TF DBDs in human

Table shows the 31 main structural classes of DBDs present in human, showing also for each of the classes the number of TFs containing the DBD and its defining Interpro identifier. Interpro is a protein family and domain classification database of the European Bioinformatics Institute (EBI).

The number of DNA binding specificities that a structural family has evolved to recognize is also highly variable, and there is very high degeneracy within many of the TF structural classes, such as Homeodomain, bHLH, bZIP, Forkhead and ETS, meaning that many TFs within a class binds to highly similar sites[72,73,81,82], while other classes, in particular the znfC2H2 TFs can recognize a much greater variety of different target sites[43,72]. The degenerate specificity of some TFs is explained by the fact that many of them are expressed in a cell type specific manner. Thus, the functional differences are not in the TF itself but in its regulatory context, the availability of chromatin and the entire ensemble of TFs that are expressed in the same cell type. This is however unlikely to be the only explanation for the large number of TFs with highly similar specificities, as they are also often expressed in the same cells or in cells that follow each other in the developmental lineage. A good example of this kind of transitional change in the expression of similar TFs are the regulatory programs driving the development of the organism through the anterior homeodomain type of TFs that recognise highly similar and simple YMATTA consensus sequences[83,84].

In the following chapters the focus will be on briefly describing the structural classes of human TFs.



Figures 4a-t | "TF-family name"; "Name of the shown TF"; PDB: "Protein Data Bank identifier"
General layout of the following figures 4a to 4t. The figure title gives the the name of the structural class, followed by the name of the representative TF shown and the protein data bank (PDB) identifier. Figure panels show the crystal- or NMR- structures for the TF shown as cartoon models from two angles and as a space-filling model. A PWM based logo for either the shown or a paralogous TF is also shown. All PWMs are derived from data from Article II, except for situations where there were no available HT-SELEX models for the indicated or paralogous TF, in which case the protein contacted part of the crystal structures DNA is shown instead. DNA is displayed with the same colors in all figures, while the protein color varies. All of the structures have been visualized using UCSC Chimera[31].

1.5.1.1 znfC2H2

C2H2 zinc fingers are the largest family of TFs with 745 predicted genes in human. This protein fold is very versatile. Apart from sequence specific DNA binding in many prominent TFs, these domains are also used to perform protein-protein, protein-RNA, protein single-stranded DNA and even protein-lipid interactions[85], thus it is very likely that some of the znf C2H2 proteins are in fact not TFs at all.

The name of this protein family is derived from the protein fold, where the peptide chain is folded into a compact structure through coordination of a zinc ion by two cysteine and two histidine residues. Unlike with other typical TFs, these domains are used in a modular fashion. There are typically from few to tens of these domains in a given TF, and domains that are ordered into tightly packed arrays are most often the ones that mediate the protein-DNA interaction. This modularity is required because an individual znf domain has only small affinity and can recognize only 3-4 bases of DNA[86].

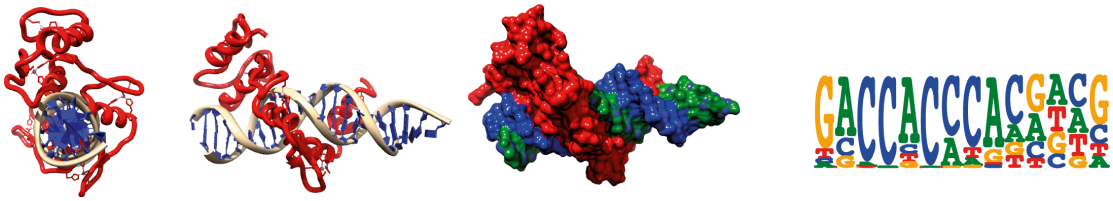


Figure 4a | znf-C2H2; GLI2; PDB:2GLI
GLI2 has 5 ZNF C2H2 domains, but only four are binding the DNA in this structure.

Before the recent findings that showed that most of these TFs (znf-C2H2 proteins with KRAB, SCAN or BTB domains) are used to silence parasitic genomic elements[22,43] they could have been described as gene regulation's dark matter, as most of them did not have proved functions, and many even lacked evidence for their existence on protein level. Consistent with functions related to "molecular immunity", this protein group has had one of the highest rates of evolution in animal kingdom and between mammalian species there are large differences in the znf-repertoires[87].

1.5.1.2 Homeodomains

Homeodomain factors comprise the second largest family with approximately 243 TFs. Homeodomains are composed of 60 amino acids and fold into a compact helix-turn-helix type of protein structure with three α-helices, one of which is the recognition helix that can contact bases in the major groove. Besides the contacts made using the recognition helix, homeodomains also contact several backbone positions, and the N-terminal "tail" inserts commonly into the minor groove of the DNA. Many of the homeodomains have important developmental roles, for example the classical homeodomains, HOX[A/B/C/D][1-13] that are used to define the anterior-posterior axis of the animal bodies[84].

The homeodomain class is large but recognizes relatively small number of specificities. For example, over half of them bind highly similar sites with YMATTA-consensus[72,82]. Most of the homeodomain TFs have only this one DBD, but in some cases it occurs in conjunction with one or more of DBDs from other classes. For example all TFs of the POU and CUT types have a homeodomain in addition to the DBDs that have given these proteins their names[84].

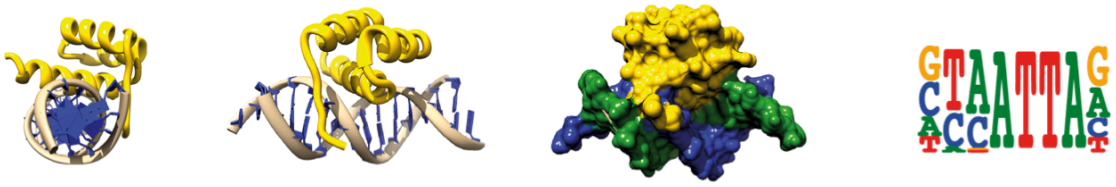


Figure 4b | Homeodomain; PDX1; PDB:2H1K

Besides the canonical YMATTA binding homeodomains there are several subgroups of homeodomains that recognize other specificities, for example, the posterior homeodomains, HOX9-HOX13 recognize a NNNTAAAA-type of sequence, PITX-class recognizes GGATTA and NK-class homeodomains recognize CACTTA consensus[82].

A special subfamily of the Homeodomains are the TALE factors, which is an ancient subgroup that existed even before the divergence of the three major kingdoms of multicellular life, the plants, fungi and animals[84,88]. This class recognizes sites that are very different from those bound by other homeodomains and there are four distinct types of them: MEIS (TGACAG), PBX (ATCA), IRX (ACAT) and MKX (TACA)[89]. TFs of this group are well known to bind cooperatively with each other or together with regular homeodomain proteins[90-92].

1.5.1.3 Basic helix-loop-helix (bHLH)

bHLH is the third largest group of TFs with 115 TFs in human. As the name suggests, the bHLH protein domain is composed of two alpha helices that are connected to each other by a peptide loop. The shorter N-terminal helix inserts into the major groove of the DNA and makes the base-specific contacts, while the loop region adds to the binding by forming contacts with the DNA backbone. The C-terminal helix is used as a protein-protein interaction interface for contacting other bHLH proteins. DNA recognition of these TFs is in all known cases based on homo- or heterodimerization with other members of the same group. bHLH family has many prominent members such as MAX and MYC, which function in the control of the cell cycle, as well as CLOCK, which functions in the control of the circadian rhythm of the cells[93].

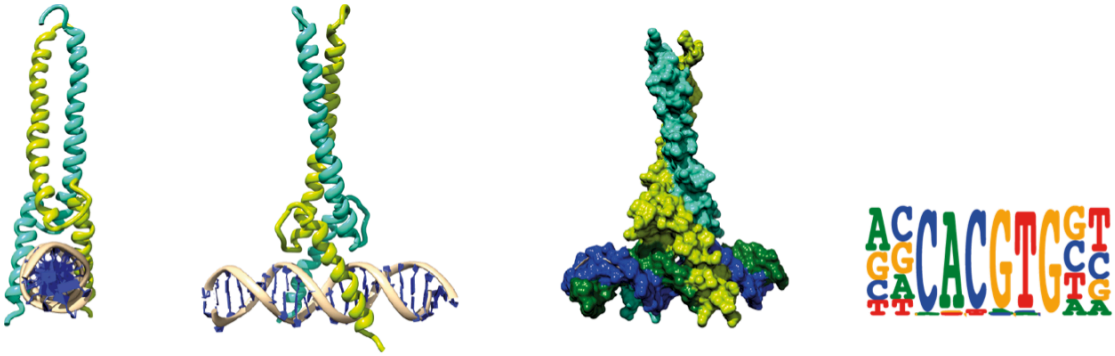


Figure 4c | bHLH; MAX; PDB:1AN2

The dimeric nature of these TFs is mirrored by the sites they recognize, which tend to have a fully or partially symmetric six bases long core sequence. While the most common binding specificity of a half-site is CAC, certain members of the family have evolved to recognize other half sites such as CAT, CAG or CGC. Besides the core sequences, some bHLH TFs have evolved additional specificity also for the bases that flank the core sequence[93].

1.5.1.4 Basic leucine zipper (bZIP)

bZIP is the fourth largest TF family with 56 TFs. As with the HLH proteins, most of these TFs need to bind as homo- or heterodimeric pairs with other members of the same structural class in order to bind the DNA. Exceptions from this rule are the members of the MAF-subfamily, which can also bind as monomers[94]. The bZIP fold is arguably the simplest DNA binding domain, as it is composed of a single alpha helix that performs both the DNA-binding and dimerization functions. Each bZIP TF has preferences for forming pairs with certain partners and these can be predicted based on the amino acid sequences of the leucine zipper regions[95]. The pairing preferences of the leucine zipper regions have also been determined experimentally and the results are mostly in good agreement with the predictions[96].

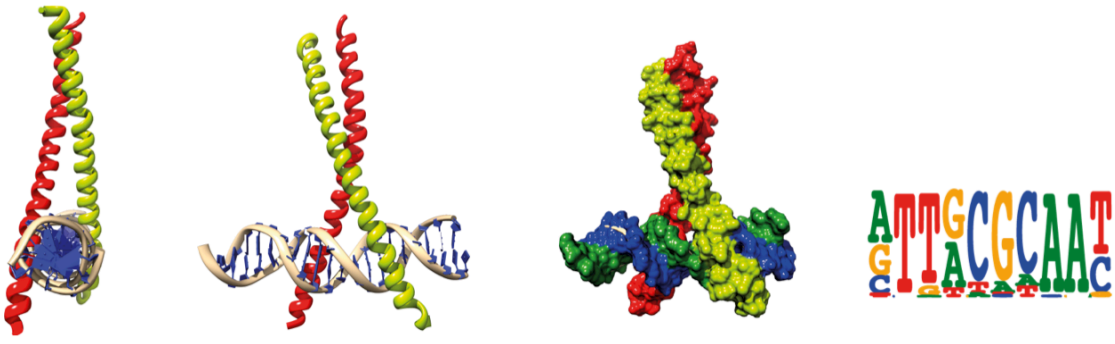


Figure 4d | bZIP; CEBPA; PDB:1NWQ

bZIP proteins have evolved a wider range of target specificities than the bHLH factors, and based on the half-sites that they recognize these TFs can be divided into several subclasses: CRE (ATGAC); CREB3 (RTGCCAC); CEBP (RTTGC) (See Fig. above); MAF (TGCTGA) and TEF (RTTAC).

1.5.1.5 Forkhead

Forkhead box (FOX) is a fairly large structural class with 49 TFs. Interestingly, the structure of these TFs, a winged helix-turn-helix protein fold, is similar to the H1 linker histones suggesting that these proteins share a common origin. This similarity is also reflected in the function of these TFs, as like the linker histones, many of these TFs can bind to DNA that is packed into a nucleosome [38,40]. Many FOX TFs function in development and consistently the knock-out mouse models die at embryonic stage or just after the birth[97]. FOX TFs can be controlled through an impressive number of post-translational modifications, as an example, FOXO1 has been shown to be able to be phosphorylated on six serines by several different kinases and is also acetylated at two lysine positions by p300/CBP, enzymes that are better known as histone acetyltransferases[98].

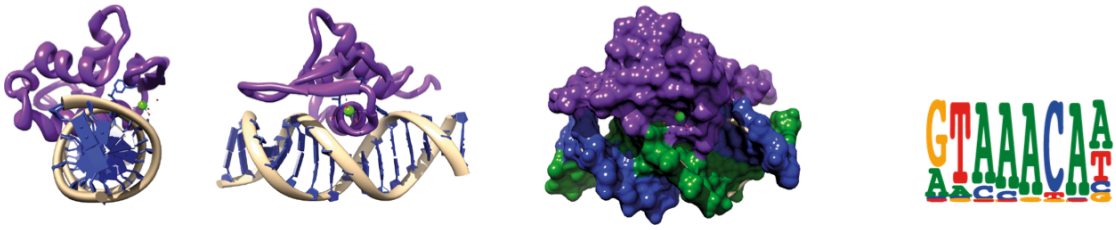


Figure 4e | Forkhead; FOXO1; PDB:3CO6

The canonical recognition site of the forkheads has an RTMAAYA-consensus, but these TFs have also evolved to bind multiple secondary target sequences such as: GACGC[99,100]; ACGGACACAAT and TTTCCCCACAC (Study II.[72]).

1.5.1.6 Nuclear receptors

There are 46 nuclear receptor TFs in human and this class is one of the most extensively studied TF families due to the clinical relevance of some of its members in e.g. breast and prostate cancers. Almost all of the nuclear receptors binds the DNA as homo- and heterodimeric complexes, although some members are also capable of binding to monomeric targets. Similar to znfC2H2 TFs, the protein fold of the nuclear receptors is also dependent on zinc ions, but is structurally very different[101].

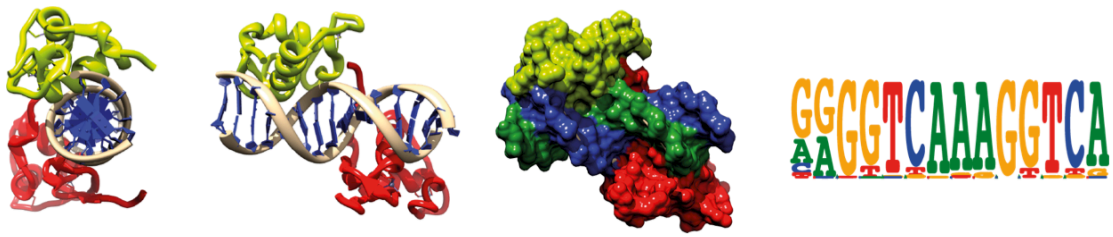


Figure 4f | Nuclear receptor; RARA:RXRA; PDB:1DSZ
The PWM is for RXRA homodimer binding the same spacing.

Nuclear receptors have evolved to recognize several types of sites with the following target consensus sequences recognized by the individual proteins: RGGTCA, which is the most common site recognized for example by estrogen receptor (ESR1) and retinoic acid receptors (RAR[A,B,G]); TCAAGGTCA, which is bound by the three estrogen receptor related TFs (ESRR); AAAGGTCA, which is bound by the NR4 subfamily; RRGWACA, which is recognized by the three members of the NR3 subfamily, one of which is the androgen receptor; RGTCCA which is bound by the two HNF4 TFs; AAGTCA, which is bound by NR2E1 and RRGTTCA, which is bound by the vitamin-D receptor. Besides their primary specificities, nuclear receptors have also diverged in terms of orientation and spacing preferences between their dimeric target sites, for example, while ESR and RAR TFs recognize the same monomeric consensus, the target sites are in different orientations; while ESR TFs bind a tail-to-tail orientation of the sites (NRGGTCANNNTGACCYN), the RAR binds to them in a direct repeat orientation (RGGTCAAAAGGTCA)[101](**Study II**[72]).

1.5.1.7 ETS

There are 27 ETS TFs in human, and the family is structurally related to FOXes as they also recognize the DNA with a winged helix-turn-helix type of domain. Many of the ETS TFs have prominent roles in basal cellular functions, their target sites occur commonly on promoter regions near the TSS, and many of them are connected to particular forms of cancers[102].

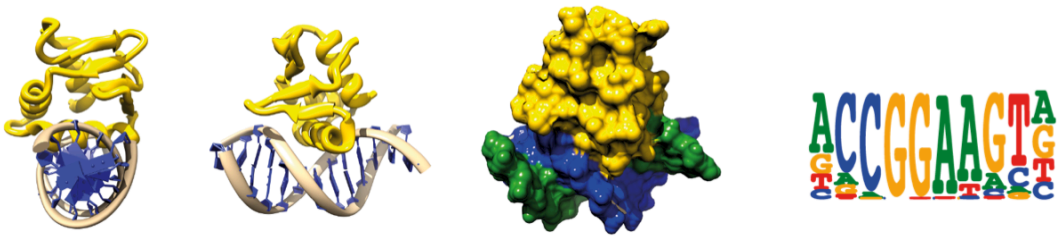


Figure 4g | ETS; ELK1; PDB:1DUX

ETS TF:s can be divided into four subclasses based on their DNA binding-specificity: ACCGGAAGN (Class I), ACCCGGAAGTN (Class II), AAAGVGGAAGTN (Class III) and ACCCGGATNN (Class IV)[81]. Besides binding as monomers, many ETS TFs can also bind as homo- or heterodimeric complexes. For example, ETS1 can bind DNA as homodimeric complex[103,104], and ETS TFs have been found to form heterodimeric complexes with partner-TFs from e.g. the FOX[105], RUNX[106] and PAX- families[107].

1.5.1.8 HMG

There are 54 High mobility group (HMG) TFs that can be divided into three categories, two of which, the SOX (20 TFs) and TCF7-like (4 TFs) types can bind DNA in a sequence specific manner, while based on both the earlier evidence and our HT-SELEX experiments, the remaining 30 HMG proteins either do not bind DNA or bind it in a non-specific manner, and thus are unlikely to be actual TFs. The most famous TF of the SOX family is the Y-chromosome located SRY, the main sex determining gene in the mammals[108].

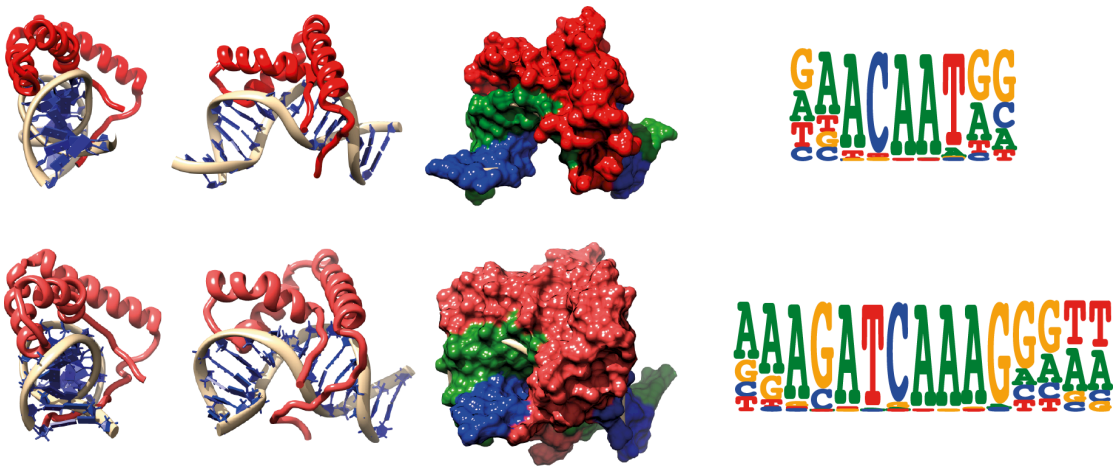


Figure 4h | *top* HMG-SOX; SOX17; PDB:3F27 | *bottom* HMG-TCF7; LEF1; PDB:2LEF

Unlike most of the other TFs, the majority of the contacts made by the HMG DBD target the minor groove of the DNA. Binding of the HMG typically causes very large effects to DNA shape. The SOX subgroup of the HMG TFs binds variant sequences of an AACAAT-consensus, while the four TCF7-like TFs bind a highly distinct AAAGATCAAAGGRWW consensus[108].

1.5.1.9 T-box

The T-box family of TFs has 17 members, many of which have been shown to function in extra-embryonic tissues such as placenta, in early development during mesoderm formation and also in later developmental contexts such as tissue morphogenesis. The developmental roles of T-box TFs are highlighted by the fact that at least ten of the T-box TFs cause embryonic or neonatal death in knockout mice, and many of them are connected to different developmental disorders[109]. The DBD of T-box factors is large and bulky, offering broad surfaces for protein-protein interaction, and it contacts both major- and minor grooves of the DNA. The recognition mechanism is however mostly indirect, as the DBD contacts the bases on the major groove with only a single hydrogen bond and from the minor groove side by phenylalanines that insert to minor groove of the DNA, while the rest of the contacts target the backbone positions of the DNA[110].

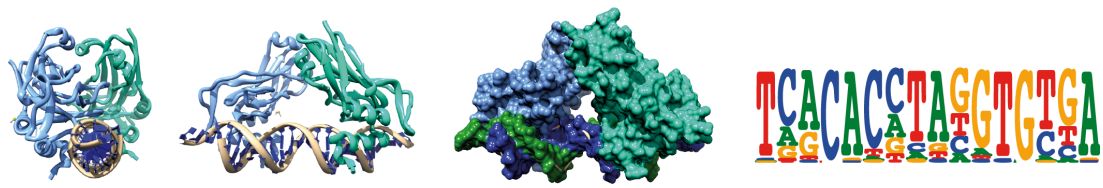


Figure 4i | T-box; TBX3; PDB:1H6F
PWM is for paralog "T" that binds the same spacing as TBX3.

All T-box TFs recognize very similar individual target sites with AGGTGTGA-consensus and they can bind this site both as monomers and homodimers. Homodimeric binding properties are different depending of the T-box paralog, at least three of the clawed frog T-box TFs have been shown to have preference for different spacing and orientation configurations[111]. In **Study II** the same was observed for many human T-box TFs[72].

1.5.1.10 POU

There are 16 TFs with a POU domain, and as described in the homeodomain chapter, all of the POU TFs contain also a homeodomain. The two protein domains are connected to each other via flexible linker regions that give these TFs the ability to recognize a large sequence space by allowing the two domains to connect to half-sites that are in different spacing and orientation combinations[112].

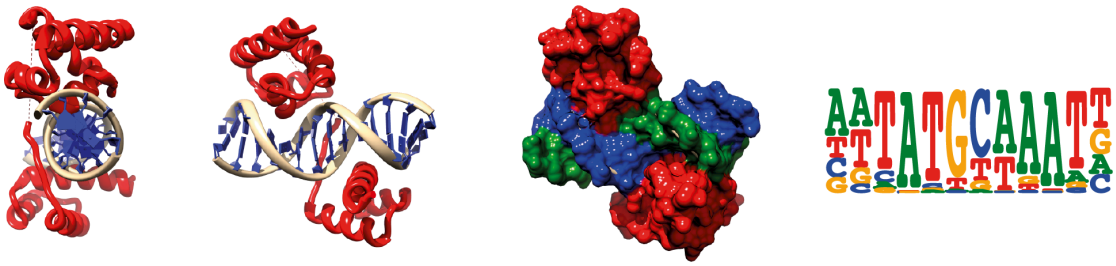


Figure 4j | POU; POU2F1; PDB:1OCT

Most POU TFs can recognize the canonical TATGCAAAT target specificity, besides which they can also recognize similar sites where there is a longer distance between the target sites of the POU-specific- and homeodomains and additionally they can also recognize target sites as homodimeric complexes. Each POU TF can typically recognize a couple of different kinds of target sites. The preference that a POU-TF has for certain types of sites is at least to some extent determined by the length and sequence of the linker region that ties the two domains together, for example in the POU5F1 crystal structure, part of its linker is folded into an alpha-helix[113], while in POU2F1 the linker is fully unstructured[114].

POU TFs are also known to make heterodimeric complexes with several different partners. Many cooperative pairs are formed between different combinations of POU and SOX TFs and depending on the pair they recognize, differently spaced POU:SOX composite motifs[115]. Interestingly the ubiquitously expressed POU2F1 has been also hijacked by Adeno- and

Herpes simplex- viruses where it heterodimerizes with viral proteins to control expression of the viral genes[112].

1.5.1.11 GATA

A GATA domain is found in 15 proteins, and at least six of them (GATA[1-6]) are proven sequence specific TFs. Similarly to FOX TFs, the GATA-family members can function as pioneer factors that can bind to their target sites even when the sequences that contain the sites are loaded with nucleosomes[38].

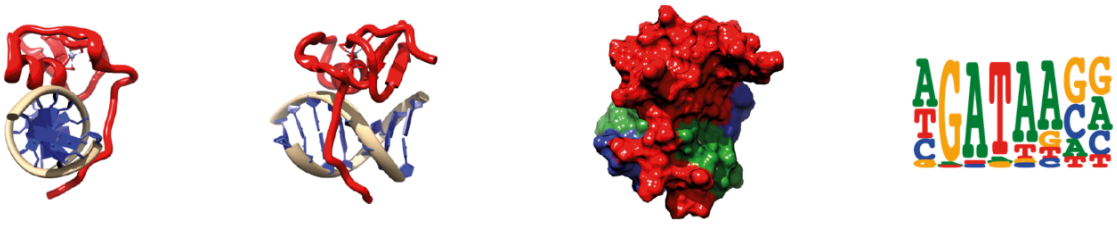


Figure 4k | GATA; GATA1; PDB:1GAT
The PWM is for the paralog GATA3, and the structure has only the N-terminal GATA-domain.

GATA[1-6] TFs have two zinc coordinated DBDs that are connected to each other by a flexible linker sequence, the C-terminal of which binds to the canonical GATAA-target site in DNA, whereas the N-terminal domain binds to GATC-sequences if present in composite sites[116]. Alternatively, the N-terminal DBD can serve as a protein-protein interaction interface with its binding partner ZFPM1 (also known as "Friend of Gata"(FOG1))[49].

1.5.1.12 E2F

There are 11 TFs with E2F domains that can be divided into two subclasses, E2F and DP domains. Six of the TFs, (E2F[1-6]) contain just an E2F domain, three (TFDP[1-3]) contain just a DP domain and two of them (E2F7 and E2F8) contain two E2F-like domains[117].

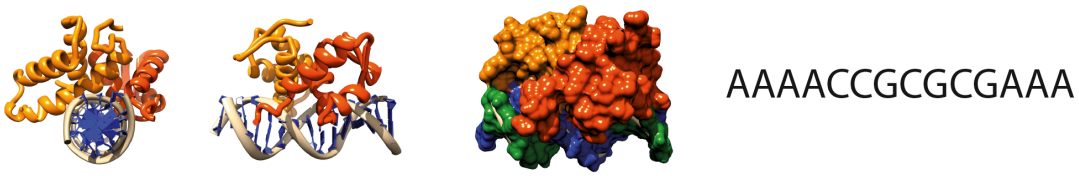


Figure 4i | E2F; E2F4(red): TFDP1(orange); PDB:1CF7
No E2F proteins were tested as heterodimers in our HT-SELEX experiments, so the sequence in the crystal structure is shown instead.

E2F[1-6] TFs bind DNA as heterodimers together with the TFDP[1-3], but it is still somewhat unclear which DNA-sequences are recognized by these complexes. For example, many older studies such as a SELEX study of multiple combinations of E2F and DP TFs[118] and a study analyzing E2F7[117], suggest divergence of the specificities of E2F-DP complexes.

However the new data, such as *in vivo* ChIP-seq experiment derived motifs for E2F[1,4,6,7] in the Homer database[119] and our HT-SELEX data for E2F[7,8], show that all of these TFs prefer very similar TTTCCCGCCAAA-like consensus. This issue still merits further research, as we have not for example tested the heterodimeric E2F combinations at all.

1.5.1.13 RHD

The Rell homology domain (RHD) class has 10 TFs that can be divided into two further branches, the NFκB and NFAT subfamilies with 5 members in both of them.

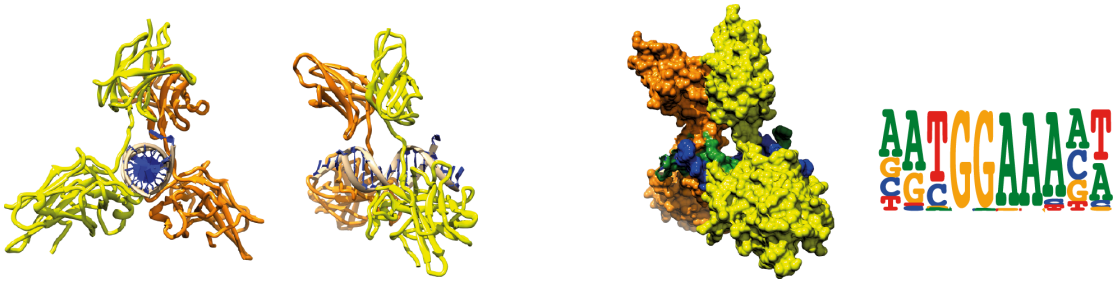


Figure 4j | RHD; NFATC2; PDB:1P7H

Members of the NFκB subfamily bind DNA as homo- or heterodimers with other members of the group recognizing a GGGGAWTCCCC consensus, while the NFAT TFs recognize a ARYGGAAANW consensus both as in mono- and homodimeric configurations[120].

1.5.1.14 Paired box (PAX)

There are 9 TFs with a PAX domain, four of which, PAX[3,4,6,7], contain also a functional homeodomain (IPR001356). The PAX domain is composed of two folded domains that are connected by a tethering region, but unlike in many other TFs with this type of layout the tethering region is an active participant in the DNA-binding which inserts fully into the minor groove of DNA[121].

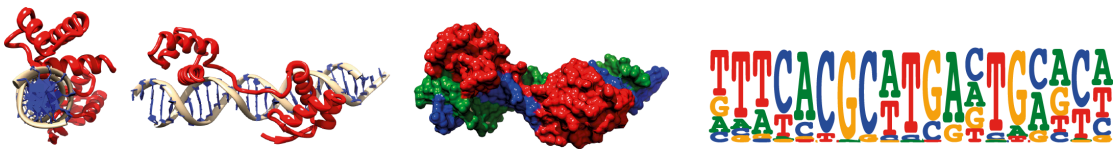


Figure 4k | PAX; PAX6; PDB:6PAX

In SELEX experiments, the PAX TFs that contain only their eponymous domain (PAX[1,2,5]) bound to a GTCACGCWTSRNTG consensus, while the PAX TFs that contain also the homeodomain can bind to more variable sites. PAX6 binds a similar site, except for a putatively homeodomain derived TTT-sequence instead of the GT (PWM is shown in the figure above, although the structure contains only the PAX domain of the TF). The HT-SELEX experiments performed using PAX[3,4,7] bound only homeodomain-type sequences

(TAATYGATTA for PAX[3,7] and general YMATTA for PAX4). Since the data is based on proteins expressed in human kidney derived cells this could for example reflect their post-translational modification state.

1.5.1.15 IRF

The family of Interferon regulatory factors (IRF) has 9 members in human. As the name implies, these TFs are strongly connected to interferon mediated antiviral responses of the cells, but they are also used in other mostly immunity related gene regulatory contexts[122].

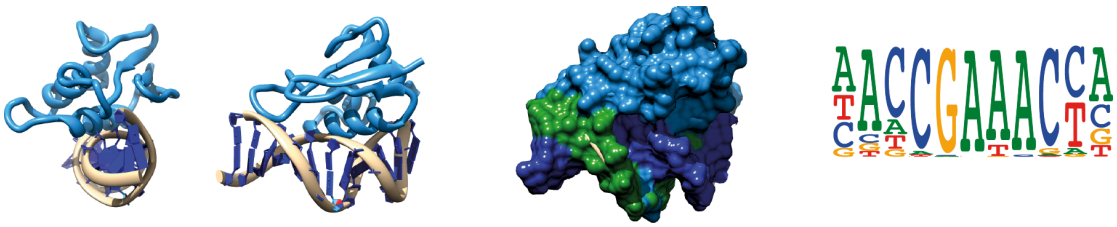


Figure 4l | IRF; IRF1; PDB:1IF1
The protein has been crystallized with a DNA hairpin; the PWM is for the paralog IRF5.

IRF TFs bind genomic target sites with often multiple copies of GAAANN consensus sequences. In the HT-SELEX we observed that the binding consensus was typically ACCGAAACYA, although some of the paralogs can also recognize secondary sites where the CCG sequence is replaced by GTG[122]. The strong tendency to form multimers was evident in our experiments performed in Study II. Based on the available crystal structures, the cooperativity of the multimeric target site binding IRF DBDs is mostly gained using close DNA allostery rather than through usage of protein-protein interactions, as in these structures there is little if any contact between the TFs[123-125].

1.5.1.16 SMAD

In human, there are 8 genes with a SMAD domain, the name of which is derived from the *C. elegans* protein SMA and the *Drosophila* protein "Mothers against decapentaplegic" (MAD)[126]. The structure of SMAD3 has been solved using X-ray crystallography, and as with the multimeric IRF structures there is very little interaction between the DBDs, suggesting that the cooperativity is gained mostly through DNA allosteric mechanisms[127].

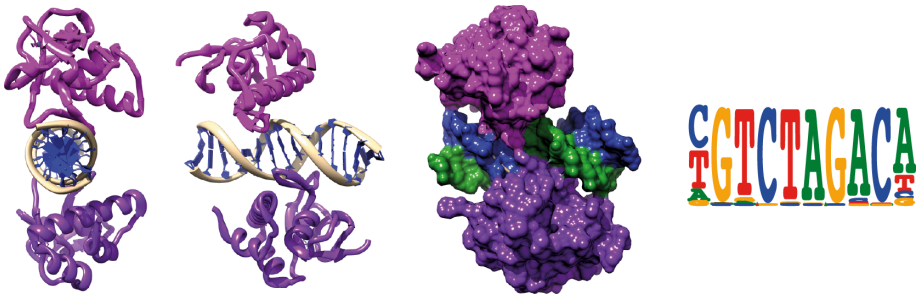


Figure 4m | SMAD; SMAD3; PDB:1OZJ

It has been suggested that SMAD TFs bind usually as heterodimeric complexes with other SMADs, in particular with the so called common mediator SMAD, SMAD4[126].

1.5.1.17 CUT

There are 7 CUT domain TFs and similarly to POU-domains this domain is always connected to a homeodomain. POU TFs can be divided into at least two clear subclasses of sequence specific TFs; ONECUT[1,2,3] TFs contain only one of both CUT- and Homeodomains, while the other type CUX[1,2], has three CUT-domains, the most C-terminal of which is directly adjacent to the single homeodomain, just as it is in the ONECUT TFs.

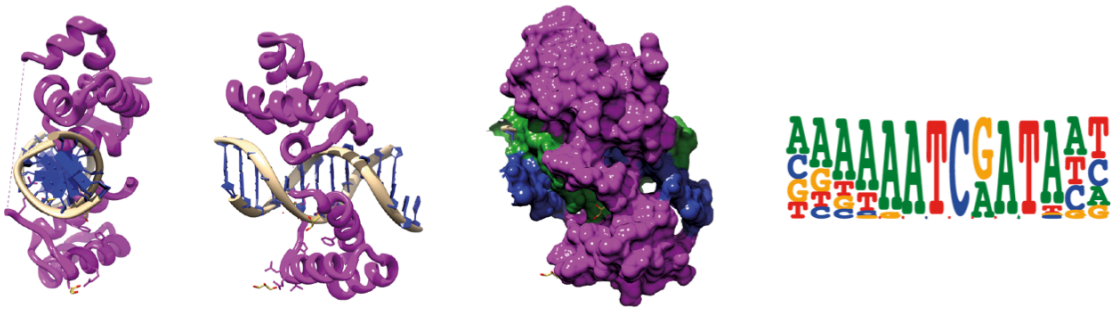


Figure 4n | CUT; ONECUT1; PDB:2D5V

Based on the X-ray structure of ONECUT1, the two DBDs of the CUT TFs are located on opposite sides of the DNA when binding to their target site[128]. Unlike with POU-domains that can bind several spacing configurations of their two sites, the two DBDs are always located within fixed distance from each other in the overlapping composite motif.

1.5.1.18 RFX

The RFX-family has 7 members in human, and as FOX and ETS TFs, its DBD is based on a winged helix-turn-helix fold. However, RFXs have evolved a different strategy for binding the DNA and recognizes different sequences[129].

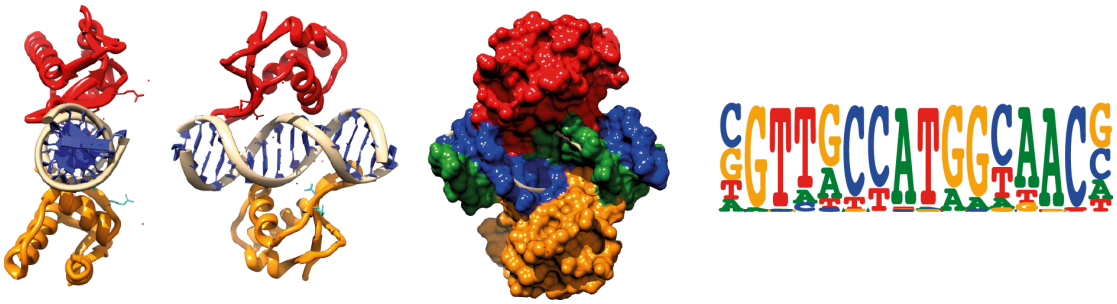


Figure 4o | RFX ;RFX1; PDB:1DP7
PWM is for the paralog RFX2.

RFX DBDs can bind DNA as monomers or homodimers[73,130]. The site bound most strongly is GTTGCCATGGCAAC, but based on both earlier SELEX-experiments[130] and our own findings from all three studies, these proteins can also recognize other binding configurations, the most relevant of which are the one and two bases shorter spacing variants of the same palindromic configuration. The structure of the RFX binding to its canonical palindromic target site has been solved using X-ray crystallography. In this structure two copies of RFX bind to opposite sides of the DNA without contacts between the DBDs, forming effectively a DNA-hamburger. This suggests that the TF uses a close DNA allostery based mechanism (see page 37) as basis for its recognition of the dimeric target sites[129].

1.5.1.19 STAT

The STAT family has 7 members. All STAT-TFs need to be activated by phosphorylation of certain tyrosine residues in order to be transported into the nucleus from their general cytoplasmic locations and to bind to DNA. STAT-TFs are used by cells in interferon signaling[131].

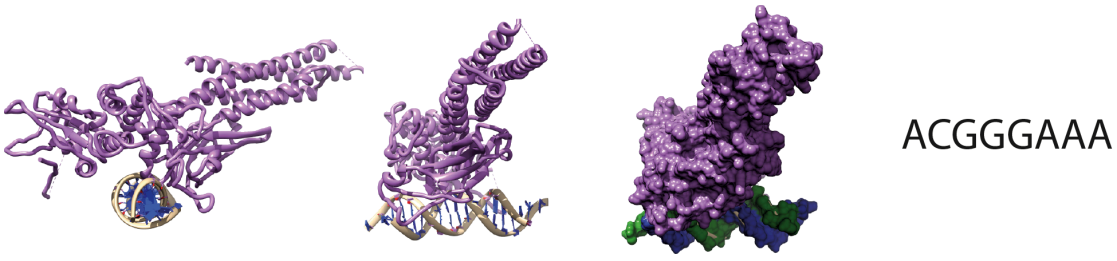


Figure 4p | STAT; STAT1; PDB:1BF5
There are no HT-SELEX models for any STAT TFs, so instead the TF contacted sequence in the crystal is shown.

All STAT TFs recognize very similar individual target sites TTC[2-4N]GAA that are bound as homodimeric configurations of the TFs, where most of the family members prefer either 3N (at least STAT[1,2A,2B]) or 4N (STAT6)[132]. Based on the crystal structure of STAT1, the STATs are also capable of binding DNA as monomers. The STAT domain is massive when compared to general DBDs[133].

1.5.1.20 MADS

MADS is an ancient TF-family with members occurring in species from all three kingdoms of life. It is a very prominent family in plants but relatively small in vertebrates with just five TFs[134].

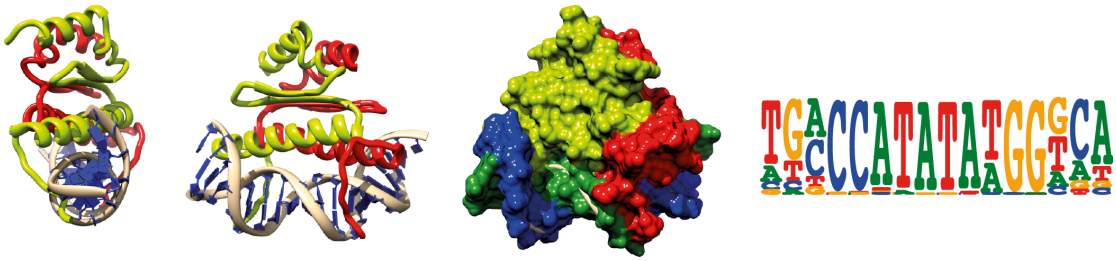


Figure 4q | MADS; SRF; PDB:1SRS

Besides the HMG TFs, this is the second class of TFs that contacts mostly the minor groove of the DNA instead of the major groove, and similarly to HMG-TFs this is also accompanied by a large shift in the shape of the DNA[135].

1.5.1.21 RUNX

The RUNX family is composed of only 3 TFs. One of the TFs of this class, RUNX1, is commonly connected to different types of leukemia, where a chromosomal fragment gets fused to a specific locations on other chromosomes leading to a fusion proteins that combine the DBD of the RUNX1 to the activator domains of other proteins. Over twenty of these kinds of gene fusions have been found, the six of which have occurred on tens to thousands of instances[136]. All RUNX TFs are connected to regulation of the cell cycle of distinct types of cells, in which they can either promote or inhibit cell division depending on either their post-translational modification state (acetylation or phosphorylation) or the cell type. Due to these roles, their silencing or activation is commonly observed in different kinds of cancers[137].

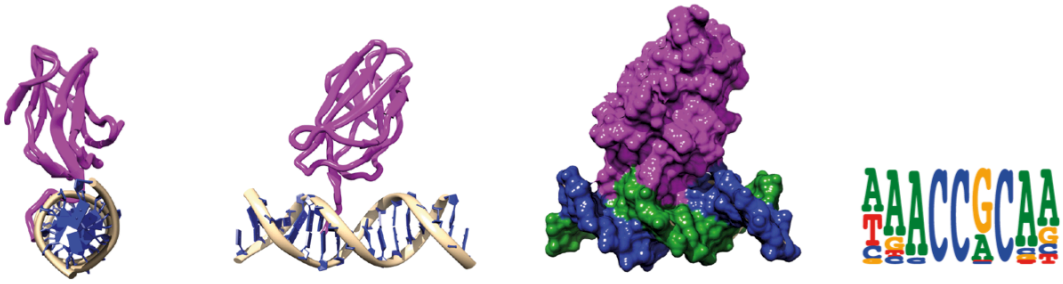


Figure 4r | RUNX; RUNX1; PDB:1HJC
PWM is for paralog RUNX2.

RUNX TFs bind sites with DNA with a short WAACCRCAA consensus sequence that can occur either as a individually or as a direct repeat that is bound by homodimer. Based on the protein structure, the footprints of these TFs are relatively small. RUNX TFs function as heterodimeric complexes with an interaction partner called CBFB. RUNX and CBFB interact through a large protein-protein interaction interface, but this is unlikely to affect the specificity as the RUNX-CBFB interface is located on the opposite site of the RUNX domain, and CBFB itself does not interact with DNA[138]. RUNX1 has been also shown to bind DNA cooperatively with ETS1 using close DNA-allostery (see page 37) as the cooperativity mechanism[106].

1.5.1.22 p53

This class contains only 3 similar TFs, but it is one of the most well known because one of the TFs, TP53, is connected to a wide variety of cancers. Typically, this TF gets either silenced or its coding sequence is lost when cells undergo malignant transformation[139,140].

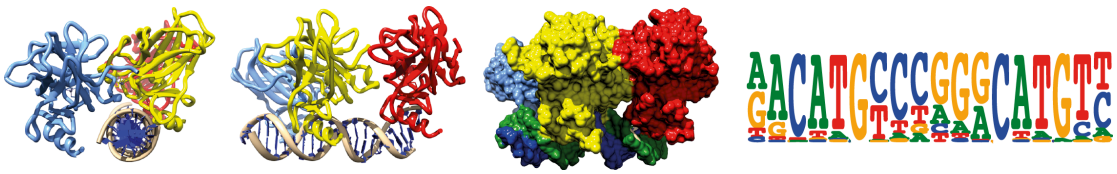


Figure 4s | p53; TP53; PDB:3F27
Configuration of the DBD is trimeric in the structure, while in reality the TP53 binds DNA as a quadrimeric complex.

All p53 TFs have a dedicated protein-protein interaction interface contained in a specific domain that causes them to form quadrimers even when occurring in free solution phase, besides this they have also a dimerization interface located in the DBD. Consistently with the layout of their multimerization interfaces, they bind DNA as 2+2 configurations recognizing duplexes of sites with RRCATGYY sequences with highest affinity for combination where the duplexes are separated by a 2bp spacer[141].

1.5.1.23 GCM

There are two GCM TFs in human. Both are connected to neural function, and the name of the class "Glial cells missing" is derived from this. Similar to C2H2 zinc fingers and nuclear receptors, this TF is also folded around coordinated zinc ions, but this is likely to be a consequence of convergent evolution, as the fold is very different from any other zinc finger domain TFs. GCM proteins are conserved broadly in metazoans, existing both in *Drosophila* and mammals[142,143].

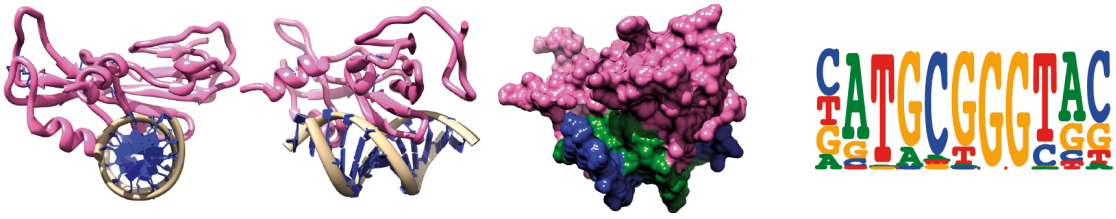


Figure 4t | GCM; GCM1; PDB:1ODH

GCM factors recognize a distinct NATGCGGGTAC consensus with a long interface to the major groove of the DNA but with no contacts to the minor groove positions or the backbone[143].

1.5.1.24 TFs families with unknown structures

The structures of DNA bound complexes have not been solved for the following TF families: Most of the heat shock factors (**HSF,** 5 TFs) have been suggested to bind DNA as trimers, although based on our findings two of them, the HSFX and HSFY, bind instead as dimers; **TEAD** (4TFs); **NFI** (4 TFs); **AP2** (5 TFs); **SAND** (9 putative TFs) at least two of which, GMEB[1,2] bind DNA in a sequence specific manner based on HT-SELEX; **CP2** (6 TFs); **DM** (7 TFs); and finally the nuclear respiratory factor **NRF** (1 TF). The single TF in the NRF-family occurs commonly in the promoter regions of genes, and due to this its target PWM is commonly found when motif detection is performed to ChIP-seq data.

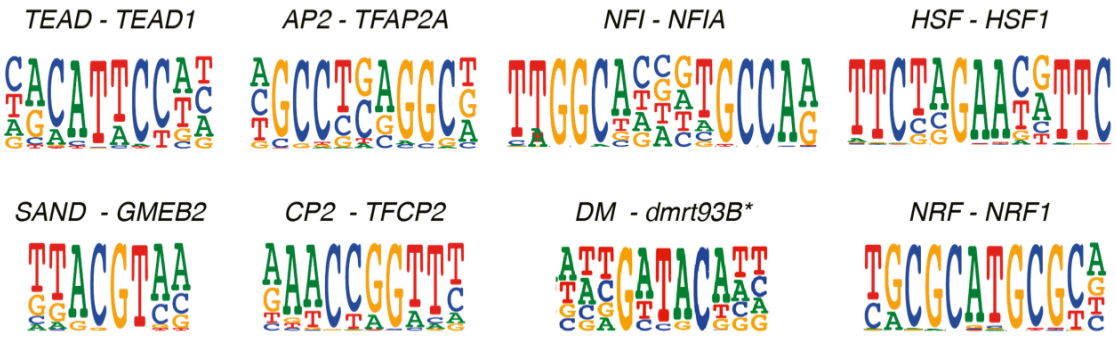


Figure 5 | PWMs for 8 TF families with unknown DBD structures

Logos for representative members of the described TF-families are shown. The logo indicated with an asterisk describes the specificity of *Drosophila melanogaster* TF (Study III), while the rest of the logos are for human TFs (Study II).

1.6 GENERAL ATTRIBUTES OF TF – DNA INTERACTIONS

1.6.1 Specific and nonspecific affinity and TF target site search process

Typical mammalian transcription factors have a large number of potential target sites that are scattered throughout the vastness of the genome, yet the cells have to respond rapidly to external and internal signals and thus the search process for target sites must be fast. There are multiple mechanisms that have been shown to be used in this process. As seen in the previous chapters, most of the regulatorily irrelevant sequences are made inaccessible for most of the TFs by wrapping them tightly into heterochromatin form. This essentially decreases the ratio of target site containing sequences to total genomic DNA and could facilitate the TF search process[144]. The search speed varies for different TFs, which has been demonstrated in fluorescence recovery after photobleaching experiments (FRAP), where the TFs that can interact with nucleosomal DNA, such as the FOXA1 pioneer factor, exhibit the slowest mobilities in the nucleus[145].

The simplest assumption, namely that the target site search process is based on diffusion followed by readout of the encountered DNA would be too slow for the cells' needs based on modelling of the nuclear conditions, even when taking only accessible DNA into account. The way the transcription factors solve this issue is to have a low nonspecific affinity to any DNA sequence in addition to the much higher affinity for their sequence specific target sites and these nonspecific contacts are employed by the TF as a way to scan the DNA sequences by sliding along it. This nonspecific affinity, typically achieved through contacts with the DNA backbone, is usually thousand times weaker (micromolar range) than the specific affinity (nanomolar range), but it is still a relatively strong interaction. In practice, the search process is probably a mixture of free diffusion within the nucleus that is then followed by episodes of two-dimensional search, the length of which is limited by DNA accessibility[146].

Advances in fluorescence microscopy based methods have made it possible to detect and analyse the TF-DNA binding of single molecules either in living bacteria[147,148], on *in vitro* arrays of single molecules[149] and even in living mammalian cells[150].

These analyses have shown for example that a specific transcription factor will stick to its specific target sites for a few minutes, while the non-specific affinity is very transient as the proteins are essentially scanning the DNA. The methods have also shown that for bacterial LacI, the average sliding distance observed in the two dimensional search process is 45 +/-1 bases. The sliding nature of the search process means also that the local chromatin environment has influence on the association rate, as any other protein that is situated within the sliding distance of the target site will function as a roadblock. The equilibrium state of the binding process is however not changed as both the association and the dissociation are affected to similar extent[148]. New evidence points towards the view that the TFs often do not recognize their target sites in a single attempt, but that it is a more of a dynamic trial and error type of process, where the TF is likely to slide over its target site on many of their encounters.

It appears that the evolution has compromised in the trade-off between fast search process and the probability that the TF will attach to its target site[148,150].

1.6.2 Mechanisms of sequence specific DNA recognition

Even though there are over twenty clearly distinct structural classes of DBDs, most of them have convergently evolved similar mechanisms to bind the DNA and also to recognize specific sequences in it. The phosphate groups located in the backbone of the DNA make it negatively charged, and thus positively charged amino acids (arginine, histidine and lysine), have affinity towards it. Moreover, the DNA backbone offers many options for formation of hydrogen bonds with both the phosphate and the sugar groups[151].

1.6.2.1 Direct readout of bases

"Direct readout" refers to mechanisms where the amino-acid residues of the proteins are forming direct and base specific contacts with the DNA bases through hydrogen bonds and hydrophobic interactions. Each of the base pairs have distinct combinations of hydrogen bond donors and acceptors in the major groove of the DNA, which can be contacted by suitably positioned amino-acid residues of a TF. Thymines and methylated cytosines have also a hydrophobic patch that can be contacted by hydrophobic amino-acid residues or by nonpolar parts of an otherwise polar amino acid. The minor groove on the other hand offers a much more limited selection of groups for recognition that is practically limited to distinguishing A:T or T:A basepairs from C:G or G:C basepairs[152].

1.6.2.2 Indirect readout through DNA shape

Indirect readout refers to the ways nature has come up with alternative strategies to circumvent the limitations of the directly recognisable positions. The shape of the DNA is dependent on its base content and the most relevant way to view this property is through analysis of the adjacent dinucleotides. In regular DNA, there are 16 possible dinucleotides and each of them has a different shape due to differences in the stacking interactions between the adjacent base pairs and the interbase hydrogen bond formation in the cases of ApA, ApT and CpG dinucleotides[153,154]. The sequence also affects the deformability of the DNA, i.e. the range of the shapes that it can accommodate when under strain. Strain can be caused by being bound by proteins such as TFs or histones or by topological means, in essence by being under- or overwound by DNA- or RNA polymerases or enzymes known as topoisomerases. A common way of TFs to recognize local DNA shape is insertion of an arginine residue into a narrowed minor groove or making backbone contacts that reach over the minor- or major grooves. Local shape of the DNA can even be interrupted to a level where the DNA melts locally, meaning that the hydrogen bonds between the paired bases are pulled apart, and base positions can get turned out of the base stacking interaction[152,155].

1.6.3 Interactions between DNA-bound TFs

It has long been suggested that the main difference between real functional motifs and randomly occurring target sites is that the functional sites are localized among other TF target

sites on regions that show evolutionary conservation. This is well supported by systematic chromatin immunoprecipitation sequencing (ChIP-seq) analyses of the binding sites of tens of TFs in the same cells indicating that TFs bind DNA commonly as densely packed, heterogenous clusters[65,156]. It is therefore very clear that TFs control gene regulation in a highly cooperative fashion. Besides cooperative interactions, TFs are also commonly competing with each other for the binding of overlapping or closely located binding sites, and all these ways of interaction are used essentially as logical operators in gene regulation.

TFs can cooperate through many different mechanisms, the most general being the passive and indirect cooperativity that is based on competition of sequence specific TFs with the almost nonspecific binding of nucleosomes. This occurs effectively because the nucleosomes' footprint of around 147 bases can easily cover several TF binding sites, and once a nucleosome is competed off, several target sites will be exposed nearby for other TFs to bind[157,158]. Other common mechanisms are protein-protein interactions, which occur commonly in many of the classical TFs. For example, most of the bZIP and all of the bHLH TFs can bind to DNA only as homo- or heterodimeric complexes with partners from the same structural class[96,159]. TFs have been also shown to form protein-protein interaction utilizing homo- and heterodimeric complexes in contexts where both of the TFs are capable of binding the DNA also as monomers, e.g. ETS1 can bind both as monomer or homodimer[103,104].

TFs can also cooperate by using DNA-allosteric mechanisms, where binding of one TF will enhance binding of another through DNA mediated means without needing direct contact between the proteins. These mechanisms can be roughly divided into two types; Direct DNA-allostery (see below for references) and oscillatory DNA-allostery[160].

Direct DNA-allostery occurs if the individual specificities of two TFs overlap. In this case, their sites are partially truncated or merged to form a composite site. When the first of the TFs binds to its target region of this composite site, it changes the shape of the DNA in a way that supports the binding of the other TF to its respective target sequence.

Intuitively, this kind of binding mechanism may seem unlikely, as one would expect a steric clash between the two TFs. However, as DNA is a helix with a full 360 degree twist every ten bases, a five base distance is enough to place the second TF on the opposite side of the helix, creating enough space for it to bind. This type of cooperativity has been observed for example in biochemical experiments that analyzed the binding strengths of POU2F1 TF to its target site in two contexts, either as an intact DBD with both its POU- and homeodomain, or in the case where the linker between these domains was severed showing that the two DBDs bind cooperatively even in the latter case and that the linker sequence strengthens the interaction merely by increasing the local effective concentration of the two domains[161]. Many crystal structures of multimers binding the same DNA support this mechanism, such as the RFX1 homodimer[129], IRF multimers[123-125], the SMAD3 homodimer[127], the FOXO1:ETS1 heterodimer(PDB id:4LG0, unpublished structure) and the structures of heteromultimeric complexes such as the enhanceosome[125] as well as the ETS1-RUNX1-CBFB complex on an ETS1-RUNX composite site[106].

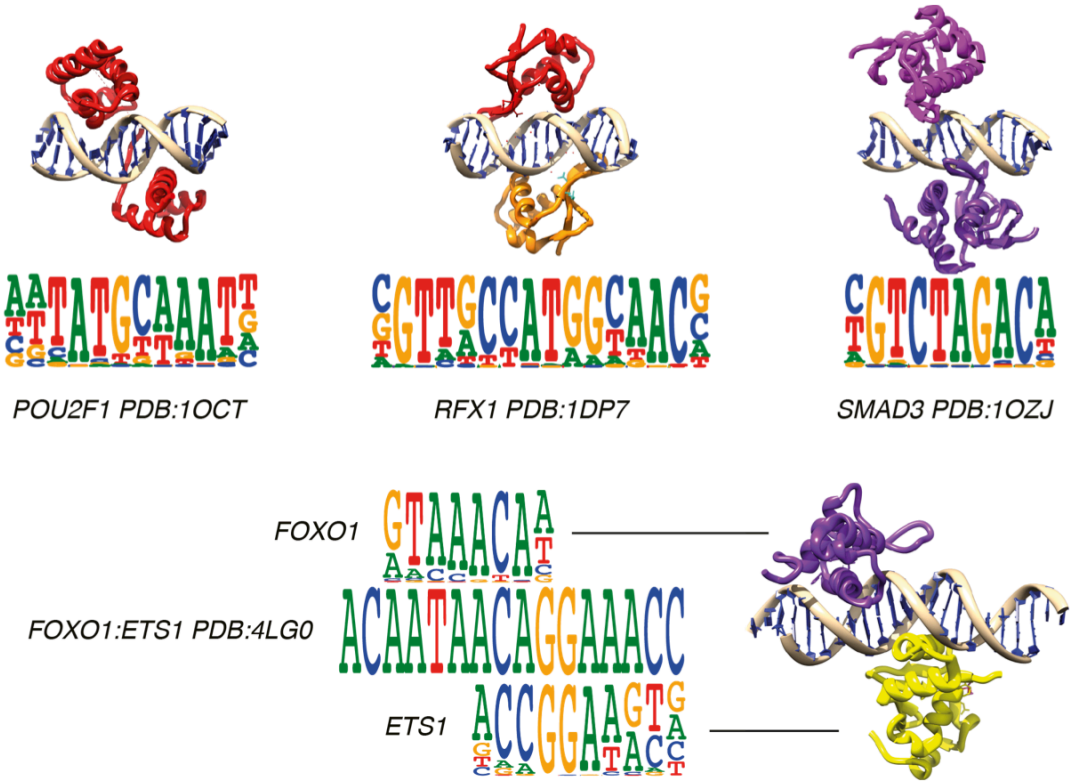


Figure 6 | DNA allosteric cooperation between TFs
Four X-ray structures where two TFs cooperate "through DNA" with either none or just minor contacts between the TFs. The PWM shown is for the same complex, except for the heterodimer example FOXO1:ETS1 (PDB id:4LG0) complex on the bottom, where the individual PWM models for FOXO1 and ETS1 are aligned to the sequence crystallized with the DBDs of these TFs.

The oscillatory DNA allostery is much more indirect and happens when pairs of TFs are binding to sites that are located in a suitable distance from each other. Typically this distance is approximately ten bases away, and the cooperativity seems to be based on the thermally excited vibrational modes of the DNA. Molecular dynamics shows that the width of the DNA's major groove changes periodicity (~10bp) due to the vibration oscillations. When an external factor, such as an DNA bound TF, distorts the major groove width of the DNA, it will energetically favor binding of a second same kind of TF to a position that is located in the suitable 10bp distance away from the other TF and disfavor its binding to sites that are half of the distance away from it. The same phenomenon applies also to asymmetric TFs, when they are binding in other orientations than the direct repeat and in instances when two different TFs are binding to DNA. Although in these cases the distances are offset when compared to the ideal 10bp homodimeric case[160].

1.6.4 Regulatory code

Detailed analyses of regulatory elements have shown large variation of their site composition, some of them have been shown to be of "billboard" type, where the sites occur as a collection of loosely packed binding sites whose orientations and spacings can be changed without major impact on target gene expression[162]. Other regulatory elements however seem to be

highly dependent on the concerted action of multiple specific TFs. A classical example of this is the interferon-β enhanceosome, where the DNA sequence is composed of a tightly packed assembly of overlapping composite binding sites and its function is easily disturbed by even small sequence changes and will drive expression efficiently only if all eight TFs are bound to it ([125], see [163] for review). As it is becoming clear that the TFs are not limited to cooperation through protein-protein interactions, it is likely that these two kinds of regulatory elements represent extremes of the possible types and in reality the regulatory elements contain local regions where the TFs bind DNA cooperatively using a wide variety of mechanisms.

2 DETERMINATION OF TRANSCRIPTION FACTOR BINDING SPECIFICITIES

2.1 TF BINDING SPECIFICITY MODELS

Transcription factor specificities can be described with many different kinds of models. TFs typically recognise a wide range of related sequences and some of them are bound with higher affinities than others. TF specificities have been described using both very simple and very sophisticated models and there is no simple answer in respect of their superiority to each other. Generally speaking, more complex models are better in describing the data they were derived from, but on the other hand simpler models are easier to evaluate and to scrutinize for artefactual features such as noise or experimental bias. Additionally, simple models are computationally less expensive.

Model quality is strongly affected by the way it has been generated and different algorithms can produce very different results from the same data due to for example the way they treat sequences with palindromic character. Model selection is also dependent on the kind of data, intended usage of the models and whether TF-specificity is approached from discriminant, probabilistic or physical angle[164]. The capability of different models to predict *in vivo* TF binding sites have been compared extensively for example in the study by Weirauch et al. 2013, where it was shown that simple position weight matrix models (PWM) perform as well as more complex models for 90% of the analysed TFs[165].

2.1.1 Consensus sequence models

The simplest way to describe the range of recognized sequences is to write it out as a string of letters A, C, G and T. Some more information can be added by using degenerate letters for positions that allow more than a single base. In the generally accepted IUPAC nomenclature these degenerate bases are marked with the letters that are shown in the **Fig. 7**[166].

2.1.2 Position weight matrix models

The position weight matrix (PWM) is a popular format for describing TF specificities. Each position of the binding site for the TF is represented by the relative preference for each of the four possible nucleotides[167]. PWM models allow a more accurate representation of the specificities than the IUPAC bases, which is not surprising as these two models are equivalent except that in the IUPAC bases the values are rounded to nearest 0.33, 0.5 or 1. A main caveat of the PWM is that it assumes that all of the positions are independent from one another. Although this generalization applies remarkably well in most of the cases[168], almost all TFs are likely to deviate from it to a small extent[169].

The PWM performs poorly in describing binding specificities of some TFs. Types of TF specificities that cannot be represented well with PWM are for example TFs that bind as dimeric complexes and allow multiple spacings and/or orientations between the binding sites. Examples of this kind of TFs include all members of the families AP2[170] and RFX[130].

The second category in which PWMs perform poorly are TFs that have strong interdependencies between their recognized bases. The independence assumption is usually more valid on the cases where the TF recognizes its bases predominantly by using direct readout mechanisms, i.e. bases are contacted directly by amino acid residues of the TF, and less valid when the recognition is based on DNA shape based readout of the DNA sequence rather than direct base contacts (see[152] and the previous chapters for details).

2.1.3 k-mer based models

The sites that TFs bind are linear sequences of DNA, and a straightforward way to model the bound sites is to describe the data as a collection of fixed length sequences that can be bound by the TF. Each of the sequences is scored based on the evidence obtained from experimental data. This model has been used mainly in describing the TF-specificity data generated from protein binding microarrays (PBM) because fairly large datasets are required to make k-mer tables of high coverage[73,82]. Such k-mer tables can, and have been also generated using our HT-SELEX data, although the PBM generated tables were better in this respect due to insufficient sequencing coverage of our published dataset[171]. The main drawback of the k-mer tables is that they are usually much shorter than the target sites (typically used as 8-mers) and thus they predict relatively high scores to partial matches to the binding sites that actually cannot be bound by the TF. This problem could be circumvented by using longer k-mers, such as 10-mers, or ideally as long as the binding specificity of the TF but no current technology can provide full coverage k-mer tables of such length. In a recent study that compared different types of models, the 8-mer based approaches did not fare better in the prediction of TF bound genomic target sites than PWM based approaches[165].

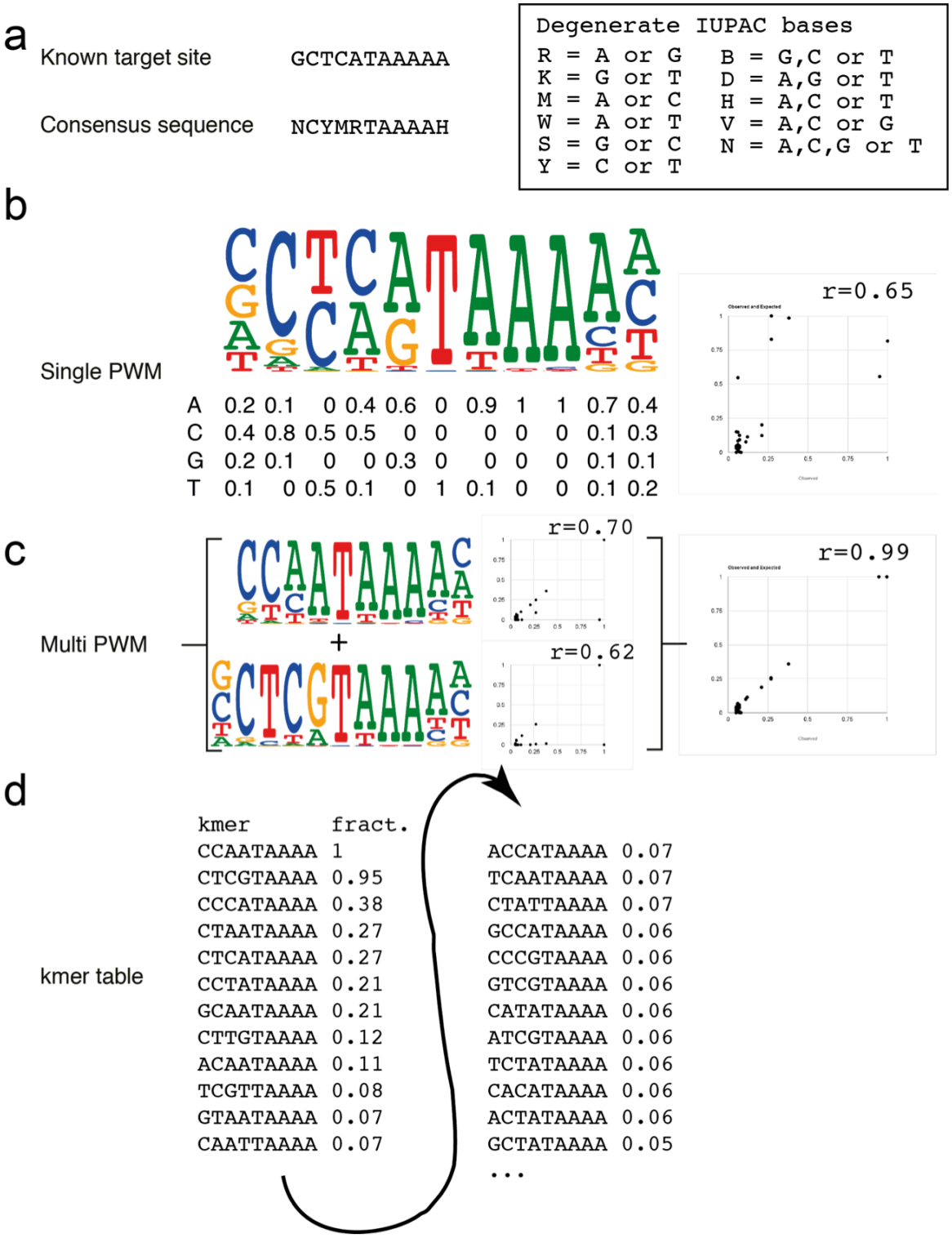


Figure 7 | TF DNA-binding specificity models

The specificity of HOXB13 is shown in six different types of models **a)** A single target site or a IUPAC degenerate consensus sequence. The box inset displays all possible degenerate IUPAC bases for the different DNA bases[166]., **b)** A single PWM shown as a logo representation (*top*) and as the actual matrix (*bottom*), **c)** Multiple PWMs, **d)** A k-mer table. Scatterplots of expected vs. observed k-mer frequencies are shown for each PWM and for the combination of the two PWMs used in the multiple PWM approach, where the reads are scored based on the PWM that gives higher score. The k-mers used for drawing the scatterplots and for calculating the Pearson correlations shown in them is the same PWM aligning subset (all 9-mers with a constant TAAAA sequence on the right side) of the k-mers that has the highest 24 of its k-mers in the ”k-mer table”. Note that while single PWM approach predicts much too large relative affinities for many of the subsequences while the multi PWM provides astonishingly good correlation.

2.1.4 Other types of models

As neither PWM nor k-mer based methods are ideal, other kinds of models have been developed, but none of them have yet gained high popularity. Some models are based on using multiple PWMs[172], or a model that connects two PWMs with linkers of defined length to make better models for TFs that bind as dimers with multiple allowed spacings[173]. Other approaches have been based on PWM-like feature based approaches that take into account the significant di- and tri-nucleotide correlations between bases[174-176]. Even more complex models have been suggested, such as Markov chains that extend the PWM (which is a zero order Markov model) by making each position dependent on the value of one or more previous positions[177,165]. A similar model, the adjacent dinucleotide matrix, was also described in Study II.

2.2 METHODS FOR SOLVING TF BINDING SPECIFICITY

TF binding can be analyzed both in *in vivo* and *in vitro*, and these should not be seen as competing but complementary approaches. *In vivo* approaches are observations of a highly complex system and as such cannot be used to explain the functions of the individual components in it. *In vivo* methodology is not the primary focus of this thesis and is discussed only briefly. *In vitro* methodology on the other hand has a central role, as the focus of all these studies is on *in vitro* analyses of TF function.

2.2.1 Methods for *in vivo* observations of TF binding

Some *in vivo* assays can provide information on the level of the entire analyzed system such as a cell line or particular tissue. Methods like DNAse-seq[178], FAIRE-seq[179] and ATAC-seq[180] are all used to determine the accessible, and thus active genomic regions by assaying either the susceptibility of the chromatin to digestion with DNAse I; the degree to which the chromatin gets crosslinked to nucleosomes when the cells are treated with formaldehyde; or is the chromatin accessible to a transposase enzyme.

More detailed information for particular chromatin associated proteins can be gained through chromatin immunoprecipitation based assays such as ChIP-chip/seq/exo[181,182], where the proteins of the chromatin are covalently crosslinked to its DNA using formaldehyde, followed by chromatin fragmentation and affinity purification of the desired protein together with the DNA fragments it is bound to. These methods are used to provide valuable snapshot-like observations of all genomic regions bound by an individual chromatin associated protein. Notably, ChIP-assays are not limited to analyses of sequence specific proteins, but have also been used to characterize genomic features such as histone modifications (reviewed for example in[14,15]).

Various methods have been also developed to map the three-dimensional organization of the genome by analysing how close genomic sequences are to each other when packed into the nucleus. Chromatin Conformation Capture (3C[183]), Circular-3C (4C[184]), Carbon-Copy 3C (5C[185]) and massively parallel sequencing adapted 3C (HiC[186]) are all variations of the same

basic idea where the DNA of the chromatin is crosslinked into its protein components followed by digestion of the DNA by a restriction enzyme and re-ligation from highly diluted solution, or within the fixed cells[59], so that the DNA strands are preferentially ligated into each other if they are covalently connected into the same complexes, and thus were located close to each other within the nucleus. Depending of the method these steps are varied to some degree and/or followed with different kind of steps for detection of the closely associated chromosomal regions, such as gel-electrophoresis, microarray-analysis or massively parallel sequencing. In a variant method called ChIA-PET these procedures are combined with additional chromatin immunoprecipitation step to gain selective coverage of interconnected regions that were associated to a certain protein[187].

2.2.2 Complementary methods for analysis of TF functions

Complementary types of methods analyze the genomic or synthetic DNA sequences for their ability to control the gene expression rates. Some assays measure the amount of mRNA molecules within the cells of interest by purifying the mRNA molecules, converting them into complementary DNA (cDNA) and then measuring the numbers of these molecules. In earlier days, and in some contexts still today, the molecule numbers were analyzed only for a few genes of interest using quantitative PCR based methods[188], but since the current DNA sequencers are capable of efficient sampling of the generated total pool of RNA, it is getting very popular to perform these kind of analyses, as they can reveal copy numbers for large amount of different transcripts. These RNA-seq assays can operate even from raw material of a single cell[189]. Besides assays that analyze the function of the genes in their native state there are systems that are used for classification of isolated gene regulatory elements. Good examples of these are the classical reporter assays, where the region of interest is cloned into a plasmid that contains a reporter gene such as a selection marker, fluorescent- or luminescent protein[190]. A modern variant of these assays is the STARR-seq assay, where the putative gene regulatory sequences are used as a library that is cloned in front of a minimal promoter and the expression rates of the regulatory elements in the library are read from the read counts of transcripts that contain the cloned fragment[191].

2.2.3 *In vitro* methods

Some of the *in vitro* methods can be used for determining TF specificities *de novo* – in essence to find the information for TFs where there was no previous available information of its specificity, while the others are useful in refining available specificity information. Despite their ripe age some historical methods are still popular, partially because of reviewers' demands, but also because some have reincarnated as modernized versions or as steps of modern assays.

One of the first methods that were used to study TF-specificities was footprinting, which is based on the idea that a DNA binding protein will protect the underlying region of DNA from being cut by agents such as DNAse I or hydroxyl radicals. On original form of the assay this protection was observed by running the partially digested and radiolabeled DNA alongside

sequencing gel so that the sequence of the protected region could be seen simply by comparing these two lanes[192-194]. The principle is still used in the form of DNAse I hypersensitivity assays, where the isolated chromatin is subjected to DNAse I digestion, which is then followed by massively parallel sequencing of the solubilized chromatin. DNAse I hypersensitivity is valuable assay as it reveals the positions of the active gene regulatory regions and the footprints of the TFs and it has been used to characterize a large number of different cell lines[45].

Filter binding assays were based on the fact that DNA does not bind to nitrocellulose membranes but protein does, and thus it provided a handy matrix for detection of protein bound DNA molecules that were usually labelled with radioactive isotopes. The method was used either to compare candidate sequences against each other, or in conjunction with *in vitro* selection assays (see below) (reviewed in[195]).

Electrophoretic mobility shift assays (EMSA) are based on the finding that electrophoretic mobility of a protein bound DNA-fragment is reduced when compared to the free DNA and this can be used as a way to separate protein bound- and free DNA fractions from each other [196-198](see[199] for a recent review). Two old *in vitro* method types, one-hybrid experiments[200] and systematic evolution of ligands by exponential enrichment (SELEX)[201,202]have also been adapted to modern high throughput methods and these will be discussed in detail in later chapters along with a third method, protein binding microarrays (PBM).

2.2.3.1 Methods for refinement of existing TF-specificity information

Once the binding specificity of a TF has been solved with e.g. one of the previously described assays, there are other types of methods that can be used to refine the information to get more accurate measurements of relative affinities to different sites. One way to evaluate the relative affinities of two different DNA molecules to a TF are competition assays. Known quantities of the two DNA oligomers are incubated with the TF followed by separation of the free and protein bound DNA and quantification of one of the two sequence species. Traditionally, the separation step was performed with filter-binding or EMSA but our laboratory has developed a more practical and accurate method for simultaneous comparison of tens of different sequences on affinity coated microwell plates using luciferase conjugated fusion proteins[203]. This microwell based competition assay can measure the relative binding affinity of any DNA-fragment against a reference DNA by competing these two species against each other. The binding affinity matrix is calculated from results of multiple parallel competition experiments, where the reference sequence (strongest target sequence) is compared against a variant in separate reactions. Models generated by competition assays are very reliable if the reference sequence is correct and if there are no significant interdependencies between the base positions of the TF binding, but when the reference sequence is incorrect or the site has interdependent base positions the competition assay will yield biased or partial models, respectively. The assay was used initially to characterize specificities of a few clinically important TFs[204] and then to systematically characterize the binding specificities of all TFs of the ETS-class[81].

2.2.3.2 Measurement of TF binding affinities

Binding affinities of TF to target sites can be measured by several methods. Best measurement of the interaction between TF and its target site is the actual physical binding constant and there are multiple assays that can be used to solve this property with various levels of accuracy, such as; EMSA or filter binding assay that can be used for rough estimation of binding affinity constants by running parallel reactions with a range of concentrations, isothermal titration calorimetry (ITC)[205] that quantifies the energy released as heat when one of the partner molecules is titrated into the solution of the other, microscale thermophoresis (MST) that can quantify the fraction of free DNA or protein in relation to their complex due to their different mobilities in microscopic thermal gradients[206] or surface plasmon resonance (SPR), which measures absorption of molecules to a thin metal sheet that is coated with the other interacting molecule, by detecting the angle where light of certain wavelength gets absorbed into the surface instead of being reflected from it. All of these methods are laborious, difficult, expensive and/or capable of running samples only in very limited throughput. A few methods have been developed that could potentially determine binding affinities in reasonable throughput, such as microfluidics based assay based on instantaneous separation of free and protein bound DNA fractions followed by fluorescence based quantification of the bound molecules (MITOMI[207]) or SPR microarray that can analyse in parallel the interaction of a single TF with tens of different target sequences[208].

2.2.3.3 Modern high throughput methods

Like many other branches of the biosciences the analysis of TF binding has recently undergone two technological revolutions, first through the development of microarrays and then by the massively parallel sequencing technologies. These inventions lead into the three main modern *in vitro* TF-specificity recognition methods, Protein Binding Microarray (PBM), high-throughput SELEX as used here (HT-SELEX) and bacterial one hybrid (B1H).

PBMs are glass slides with 40,000 printed spots of 60 bases long individual single stranded oligonucleotides that have been designed to contain all possible 10-mers at least once. This microarray is converted into double-stranded DNA by synthesis of the second strand, which is then followed by binding of an epitope tagged TF construct, washing, and fluorescence based quantification of the amount of protein bound to the individual spots[209-211]. A similar approach, cognate site identifier (CSI), differs from the PBM in the way that the sequences on the microarray form the double-stranded DNA through making a loop on the center and self-annealing. This method has been used for solving the specificities DNA binding small molecules rather than TFs[212].

In SELEX assays the DNA-binding proteins are incubated with double-stranded DNA oligos composed of random sequence that is flanked by constant amplification regions, allowing different DNA “ligands” to compete for binding to recombinant protein. After the binding step the protein bound oligomers are separated from free DNA either by EMSA or by affinity purification of the proteins. Protein bound DNA-oligomers are then multiplied by PCR,

purified and then used as a new ligand-pool for another cycle of selection. Each round of purification followed by PCR enriches sequences in a manner related to their affinity towards the ligands, i.e. a sequence which has 10-fold lower affinity to a TF than the sequence with highest affinity will be present at 10-fold lower concentration after the first cycle, and 100-fold lower concentration after the second cycle[201,202].

Bacterial one hybrid (B1H) is based on two kinds of plasmids. One plasmid type is used as a library of plasmids that is generated by cloning short pieces of randomized DNA into a plasmid position that is located just before a minimal promoter sequence. This minimal promoter is designed to be by itself able to drive just very weak transcription rate, and thus it requires additional TF-sites present in subset of the cloned random sequences to drive effective transcription. In the plasmid this promoter drives transcription of an mRNA sequence that contains two bacterial strain and growth condition specific selection marker genes. One of the selection markers codes for a bacterial protein that can be used for negative selection (kills the bacteria) and the other one can be used for positive selection (bacteria stay alive) of sequences that activate the promoter. The second plasmid type codes for the DNA binding domain of the desired protein fused to a partial bacterial RNA polymerase[213].

As an initial phase of the B1H experiment the clone library is used to transform bacteria which are then grown in counterselection conditions designed to kill the bacteria that contain plasmids with self-activating sequences that can drive expression of the promoter using the bacteria's own TFs. In the following stage, the actual selection is performed by transforming the bacteria also with the second plasmid and performing selection in a different kind of medium, where the survival of the bacteria is dependent on the promoter driven expression of the positive selection marker on the first plasmid[213].

B1H has been used primarily in Scott A. Wolfe's lab for modeling the specificities of TFs of the fruit fly *Drosophila melanogaster*. First, 84 Drosophila homeodomains[214] were modeled, followed by 35 TFs regulating Drosophila segmentation[215] and finally 129 zinc finger C2H2 TFs[216].

3 AIMS OF THE STUDY

All three studies included in this thesis aim at the generation of information about the DNA binding-specificities of TFs. From the beginning the purpose was to be able to perform the analyses on the scale of hundreds of TFs using partially automatic pipelines.

Information on TF binding specificities is required for understanding the function of the genome. Without this information, we cannot decipher the language of the gene regulatory regions that determine when, where and how much the genes are expressed. Furthermore, the blueprints of the animals are also encoded in the 10-20% of the genome that contains the gene regulatory elements and we need the TF binding-specificities to understand what makes the already sequenced 119 vertebrate species different from each other and ourselves. The specific aims were:

1) Develop a new high-throughput method for TF-binding specificity determination.
2) Apply the method to solve binding specificities of as many human TFs as possible.
3) Study the evolutionary relationships of the DNA binding specificities of TFs by comparing data from human and from the insect model organism *Drosophila melanogaster*.

4 MATERIALS AND METHODS

Method	Study
Cell culture and transfection	I,II,III
Chromatin immunoprecipitation	I,II,III
HT-SELEX	I,II,III
HT-SELEX data analysis based on autoseed program	I,II,III
HT-SELEX data analysis based on IniMotif program	I,II
In silico analysis of genomic enrichment	I,II,III
Laboratory automation	II,III
Massively parallel sequencing with Illumina Genome analyzer	I,II,III
Real time quantitative PCR	II,III
Recombinant DNA techniques	I,II,III
Recombinant protein production in *Escherichia coli*	II,III
Recombinant protein production in mammalian cells	I,II

Data deposition

All Illumina sequencing reads connected to SELEX and ChIP-seq samples were deposited into either as a supplementary file that is located on the server of the journal Genome Research (Study I) or into ENA (European Nucleotide Archive), under accession numbers SRA012198 (Study I, ChIP-seq), ERP001824 (Study II, HT-SELEX), ERP001826 (Study II, ChIP-seq) and PRJEB7373 (Study III, HT-SELEX).

5 RESULTS

5.1 STUDY I: MULTIPLEXED MASSIVELY PARALLEL SELEX FOR CHARACTERIZATION OF HUMAN TRANSCRIPTION FACTOR BINDING SPECIFICITIES

This study was about the development of our novel HT-SELEX assay, which is essentially a high throughput adaptation of the classical SELEX method. In HT-SELEX, several of the bottlenecks of the original approach have been streamlined to allow studying TF specificities on a genomic scale. Mammalian expression plasmids were designed that allow the expression of TFs as constructs with *Gaussia princeps* luciferase and a streptavidin-binding peptide epitope. This design allows easy quantification of the yields of the TF fusion proteins and their convenient immobilization and purification using a wide variety of commonly available streptavidin based affinity matrices. The design made it also possible for us to convert the assay for laboratory automatization in **Study II,** increasing the possible throughput even further. Selection ligands were designed such that they were directly compatible with multiplexed massively parallel sequencing, which makes it possible and economically feasible to run the entire assay in batches of hundreds of TF constructs.

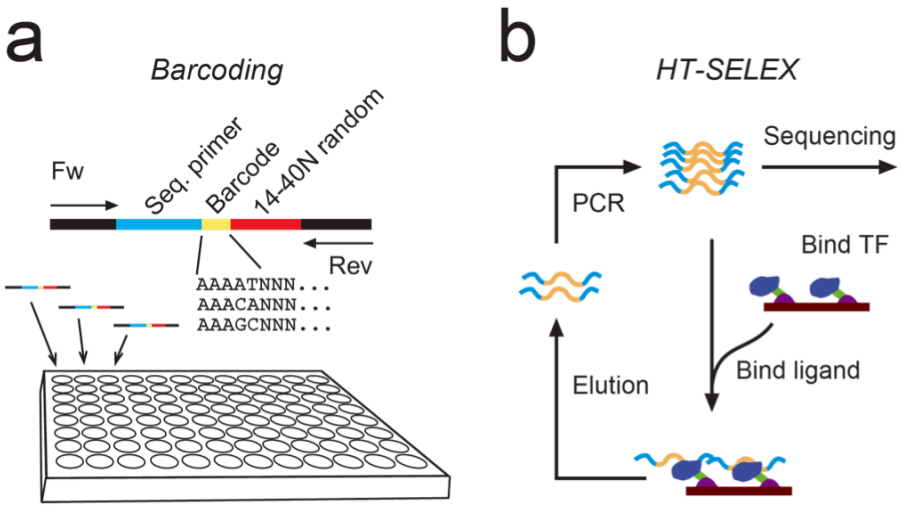


Figure 8 | HT-SELEX design

a) The selection ligands contain a barcode sequence that allows identification of the samples after multiplexed sequencing. Each well of the 96-well plate contains a ligand with a different barcode and thus represents an independent experiment. **b)** HT-SELEX is performed in a cyclical manner; the recombinant TF fusion is immobilized onto a solid support that is either the bottom of an affinity matrix coated plate or beads, followed by addition of the selection ligand. The reaction is incubated to allow the TF to find and bind to its preferred target sites from within the randomized part of the ligands. In the next step, the solid support is washed to remove unbound ligands. Protein-bound ligands are eluted from the solid support by heating to denature the TFs. Eluted ligands are then amplified with PCR to form the new pool of sequences which is either used for a further cycle of SELEX or for massively parallel sequencing. The figure is modified from[213].

The HT-SELEX assays that were run to test the assay yielded robust data for over hundred TFs, out of which we published PWM models covering 18 TFs from 14 out of the main 31 structural classes. In a successful HT-SELEX experiment, the fractions of the reads increase in each cycle until finally, if the selection would be carried out far enough, only a single or very few sequences would remain (See figure 9, below). However, in practice only four or less cycles need to be carried out to reach enrichment that is sufficient for the generation of PWM models.

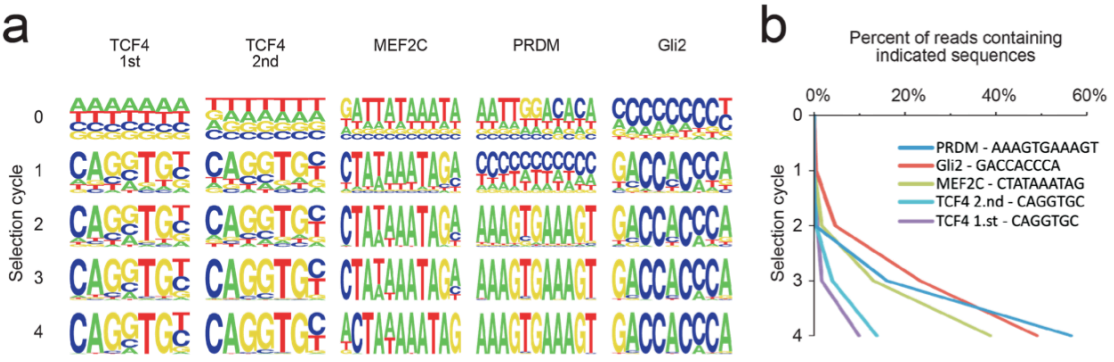


Figure 9 | Enrichment of subsequences

a) PWM models constructed for the indicated TFs and selection cycles using the same seed sequences. TCF4 was run in two replicate experiments. **b)** The graph shows the fraction of total reads that are described by the TF consensus binding sequence as a function of the selection cycle. The figure is modified from[213].

The obtained data was compared to previous findings in the literature and also with *in vivo* data obtained from ChIP-seq experiments that we carried out to validate some of the findings[73,82,130,170,217]. The produced PWM models were in very good agreement when compared against high quality models gathered from reliable sources, such as the ETS-factor-profiles evaluated independently by multiple different methods (ChIP-seq, PBM and microplate binding assay)[81]. The conclusion was that the HT-SELEX is efficient and reliable method for determination of the TF binding specificities.

The PWMs were generated in this study using a novel background corrected multinomial-1 based algorithm. The multinomial-1 approach generates the PWMs based on the seed sequence and all of the sequences that differ from it by substitution of a single base. The seed sequence is usually the most enriched sequence that covers all of the base positions that contribute significantly to the binding specificity. This is a very simple strategy of PWM generation, and follows exactly the independence assumption that is an inherent property of PWM models. As a drawback, the multinomial-1 method requires much higher numbers of sequences than the traditional alignment based methods, as it discards a lot of the enriched signal. On the other hand, the traditional alignment based methods lead to data overfitting and thus essentially any seeded sequence would yield a motif even when there is no real signal (if a sufficient number of reads are available). Additionally, the overfitting caused by the alignment based methods will make the real signal overtly stringent when compared to the physical binding affinities (See **Fig 10**).

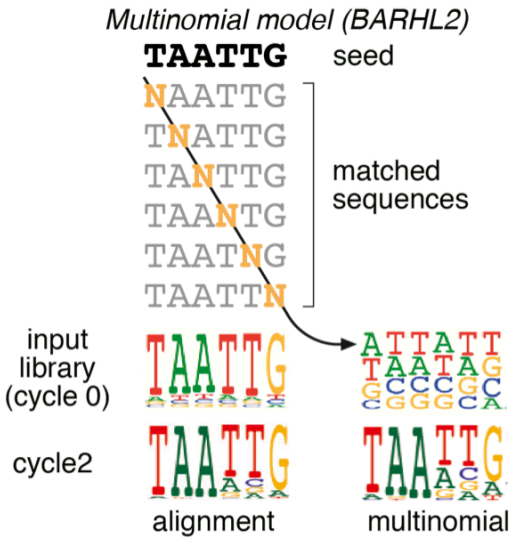


Figure 10 | Alignment vs. multinomial-1 PWM generation

Figure shows how the multinomial-1 algorithm is used to generate the PWM by counting the occurrences of each base at a given position when all other bases exactly match a seed sequence. Note that usage of simple alignment would generate an excessively stringent model, even when analyzing random sequences of the SELEX input library, while multinomial alignment does not. The figure is adapted from[72].

5.2 STUDY II: DNA-BINDING SPECIFICITIES OF HUMAN TRANSCRIPTION FACTORS

Study II applied the assay developed in the Study I for the analysis of a much larger number of TFs. We cloned collections of TFs and their DBDs from human and mouse into the pDEST40-HTSELEX (Made in **Study I**) or pETG20A[218] vectors for expression in human embryonic kidney derived 293FT cells or in *E. coli* bacterial expression system, respectively and performed the HT-SELEX with the produced fusion proteins. The analysis yielded 831 PWM models describing the binding specificities of 411 TFs representing almost all structural classes occurring in mammals.

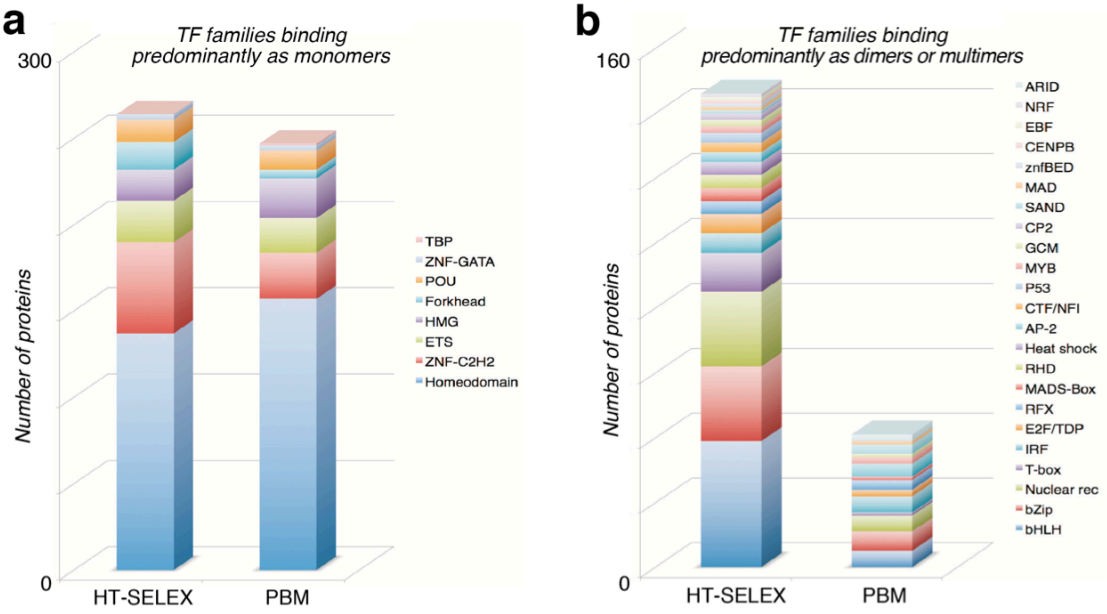


Figure 11 | Coverage of TF specificities determined using HT-SELEX and PBM.
The coverage of HT-SELEX and PBM based models for different structural classes of TFs are shown. While the coverage is similar for TFs that bind as monomers (panel a), HT-SELEX has much higher coverage for TFs that bind as multimers (panel b). The figure is adapted from[72].

Comparison of the PWM models generated using the three different types of clones (human DBDs, human full length TFs, and mouse DBDs) showed that the PWMs were essentially identical when comparing either the pairs of PWMs derived using the two human clone types or when comparing the PWMs for the human TFs and their mouse orthologs. These findings support the notion that TFs are highly modular proteins, and that evolutionary conservation of TF specificities in mammals is very high.

As seen in **Study I**, many of the TFs that were thought to bind as monomers bound DNA also as cooperative complexes of two or more proteins. This cooperative, multimeric binding revealed that TFs that bound very similar sites as monomers recognized different target sequences when binding the DNA as cooperative complexes. Some of the TFs displayed even "latent specificities" when binding as homodimers, meaning that the individual specificities of some of the paralogs, such as the ETS factors ERG and ELK1, were changed when they bound their preferred sites that were occurring in closely packed "overlapping" spacing and orientation configurations (see also[219]).

In contrast to previous work performed using PBM technology, where occurrence of nucleotides in different positions of recognized sequence was often interdependent leading to primary and secondary PWM for most of the TFs[73], our analysis showed that the majority of the TFs bind DNA in a highly position independent fashion. In most of the cases where multiple PWMs were required for description of the TF's specificities, this was due to the presence of both mono- and multimeric binding configurations, or of dimers with multiple accepted orientation and spacing combinations between the individual sites. However, we

observed some notable exceptions where the TFs required multiple PWM models even though their recognized sites did not display obviously multimeric character.

Systematic analysis of all bases in all PWMs showed that in most of the cases the base interdependencies occurred on adjacent nucleotides. This suggests that these dinucleotide preferences are based on DNA shape recognition based mechanisms. All 16 combinations of adjacent DNA basepairs have distinct shape and deformability properties, and the TFs can recognize sequences indirectly by forming shape dependent contacts to the backbone and the minor groove of the DNA. In many cases, such dinucleotide correlations were visible through enrichment of poly A or T stretches, which are often recognized by arginine residues that insert into the narrowed minor groove typical for poly A or T[152]. Some other TFs however bound sites that had more distinct types of base interdependencies. For example all four HOX13 TFs bound two sites with CTCGTAAA or CCAATAAA consensus sequences with a very strong trinucleotide correlation.

We experimented also with two novel kinds of models for description of the specificities of the TFs that cannot be described well using a single PWM. The first of these two models, the adjacent dinucleotide model, takes into account the effect of dinucleotide-level correlations between the adjacent bases. This type of model is well suited for the description of specificities of TFs like E2F3, which binds as homodimer and whose core sequence is flanked on either side by approximately four bases long stretches of A or T (Fig 12a). The second model on the other hand, the connecting matrix model, is intended for the description of TFs, such as T-box TF TBX20 that can bind as homodimers with multiple spacing and orientation combinations. In the connecting matrix model the data is represented by combining the monomeric specificity describing PWM of the TF with connecting matrix that gives rank and weight to the different spacing and orientation observations (Fig 12b).

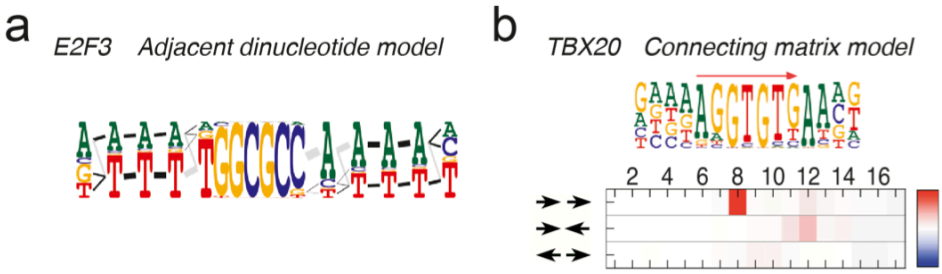


Figure 12 | Adjacent dinucleotide and connecting matrix models
a) The adjacent dinucleotide model describes TF specificity as probabilities that a base is located next to another. In the shown logo representation the overrepresented dinucleotides are indicated by black bars and gray bars represent dinucleotides that are not overrepresented when compared to expected (based on positional independence assumption). Width of the bars represents the frequency of the indicated dinucleotide. **b)** The connecting matrix model consists of a monomer PWM (monomeric site of T box TF TBX20 indicated by red arrow) and a matrix that describes spacing and orientation preferences of two such sites. The heatmap indicates that for this TF the preferred orientation and spacing configuration is when the two sites are in the same orientation and separated by an eight bases long gap. The figure is adapted from[72].

5.3 STUDY III. THE CONSERVATION OF TRANSCRIPTION FACTOR BINDING SPECIFICITIES ACROSS 600 MILLION YEARS OF BILATERIA EVOLUTION

In the previous study we had analyzed difference between mouse and human TFs and found no clear cases of changed specificities between pairs of directly orthologous TFs from these two species. This was perhaps not very surprising for the researchers who work in the field of TF specificities. After all mouse and human are close relatives when it comes to molecular evolution and the amino-acid sequences of DBD regions are very conserved between these two species.

Thus, to provide answers to the question of evolutionary conservation, we applied the HT-SELEX for analysis of binding specificities of *Drosophila melanogaster* TFs, with which our lineage shared a common ancestor over 600 million years ago. Reasons for the selection of this species was because of this long evolutionary distance and the fact that *Drosophila* is an extensively researched model organism especially in the field of developmental biology.

The clones for the *Drosophila* DBDs or full length TFs were obtained from collaborators in Eileen Furlong's and Max Deplancke's groups or ordered as synthetic genes. The TFs were produced as fusion proteins using the same *E. coli* expression based protocol as was used in Study II for a subset of the analyzed TFs. Overall we obtained PWMs for 242 TFs of *Drosophila melanogaster* describing specificities for representatives of almost all of the TF families expressed in it.

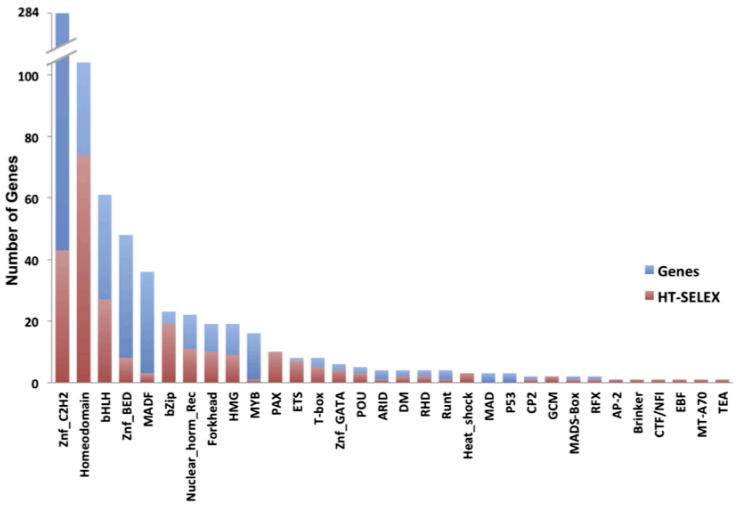


Figure 13 | Coverage of *Drosophila* TF specificities

The obtained coverage of HT-SELEX models for different structural classes of predicted *Drosophila melanogaster* TFs is shown Figure is adapted from[220].

Comparison of the fruit fly PWM dataset generated in this study against the human and mouse data derived from Study II revealed a surprisingly high level of conservation of the sequence specificities of TFs. In most cases, the fruit fly TFs bound almost identical target sites as their mammalian counterparts. Observed conservation of the binding specificities extended even to subtle dinucleotide preferences. Thus, it seems that the primary target specificities are under very strong selection pressures and have been kept remarkably constant. However, mammalian genomes code for many more sequence specific TFs (around 1400) than the fruit fly (around 400-700), and both of the lineages contain a few types of TFs that have evolved after the lineages split and that thus recognize sequences used only in the respective lineages. These include several structural classes that appear to be used in only one of the species, such as brinker TFs in *Drosophila* and IRF TFs in human. Additionally there were new subfamilies of shared structural classes that only occurred in one of the species, such as the ETS class III that has three TFs in human and has evolved different specificity in the form of an A-stack on the 5' side of the canonical ETS motif. There were even a few cases where apparently directly paralogous TFs had diverged in specificities, such as *Drosophila* CG30420 and its human orthologue ATF7. In most of the cases where TFs had evolved new specificities, they were connected to cell types that were specific to that particular animal.

In addition to the analysis of divergence of TF specificities, the article introduced several new technological innovations, such as a new distance metric "Huddinge distance" that can be used to find local maxima of overrepresented gapped DNA k-mers. The concept was implemented in two novel informatics tools, "kmerseed" and "autoseed" that allowed a much deeper analysis of sequence patterns than previously possible using existing tools. As these new methods can analyze all possible gapped or ungapped k-mers and then output the desired number of highest local maxima, they excel in sensitive detection of various enriched signals (as implemented in the program "autoseed") and distinguishing the most relevant spacings and orientations between homo- or heterodimeric target sites (in "kmerseed").

Our data supports the view that, with the exception of the highly modular family of C2H2 zinc fingers, the protein-DNA interactions are very resistant to evolutionary changes. Thus, the majority of differences in gene regulation, and thus most of the differences between two organisms are not in the machinery of transcriptional regulation but in the DNA-sequences that are read by them. Thus, analogously to the genetic code that specifies how the DNA sequence of the genes is converted to amino acid sequences of the proteins, the second genetic code that guides transcription is very similar between animals and the changes occur in the information itself and not in the way it is decoded in different organisms.

6 DISCUSSION

Before our studies, the knowledge of TF-binding specificities was very fragmented. Most of the data had been generated by hundreds of individual researchers that were using a multitude of different methods, and in almost all of these experiments the models were based on very few measurements. Because of these reasons it was very hard to evaluate the quality of the models and comparison of different data was often very likely to reflect method specific biases rather than genuine differences in the binding specificities of the TFs.

Notable exceptions to this were the previous large-scale analyses of TF specificities for several hundred mouse and *Drosophila melanogaster* TFs using PBM and B1H, respectively[73,82,214,215,216]. Additionally, our own group had analyzed the specificities of the entire ETS class using a microwell based competition assay[81].

The value of consistent datasets generated by the modern high throughput assays is shown clearly when the older models assembled from literature are compared to results from PBM, B1H, HT-SELEX or the microwell based competition assay[81,82] (**Fig 14**).

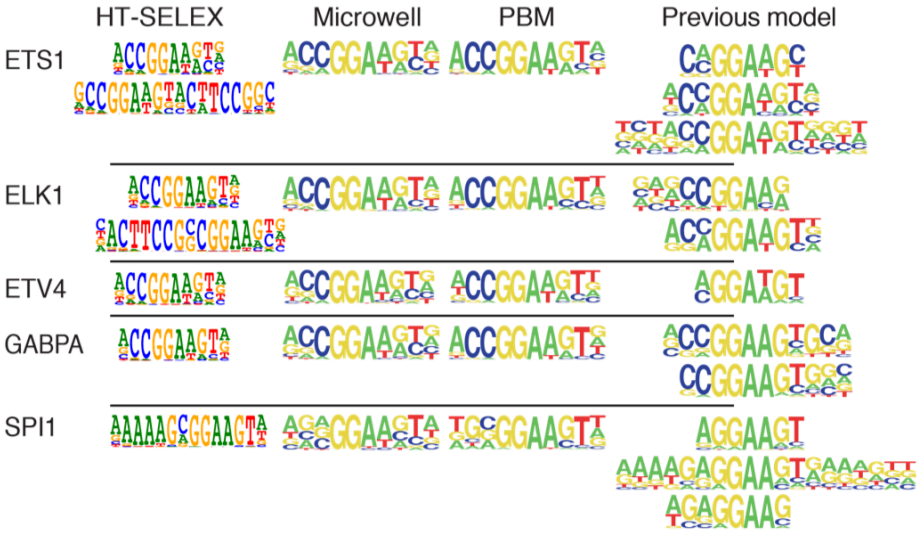


Figure 14 | Benefits of systematically generated datasets

Figure compares PWM models generated systematically using three modern methods, HT-SELEX, Microwell based competition assay and PBM to older models collected from literature into Jaspar database[217]. The four topmost TFs belong to ETS class I and bind very similar primary PWMs based on all three modern methods. The PWMs assembled from the literature are on the other hand very heterogenous and analysis based on them would lead to false assumption that the TFs have divergent primary specificities. Note also that only the HT-SELEX has captured the dimeric binding modes of ETS1 and ELK1 shown below the monomeric PWM, and that only the HT-SELEX and one of the Jaspar models (based on ChIP-seq *in vivo* data) have detected the "A-stack" of SPI1. The microwell based competition assay and universal PBM cannot detect these binding modes because the microwell based competition requires previous information about a consensus binding site and the universal PBM contains only all possible 10 base long sequences. The Figure is modified from[81] through addition of data from[72].

6.1.1.1 Benefits of the HT-SELEX over other modern methods

PBM method is limited mainly by its ability to cover only all up to 10 bases long possible sequences because the number of all possible sequences is exponentially related to the length of the DNA k-mer. Due to this inherent limitation the PBM often fails to model the specificities of the TFs that recognize long target sequences, or it ends up into generation of PWMs that describe poorly the specificity of the TF e.g. by modeling a subset of the target site (See Figure 15, *below*).

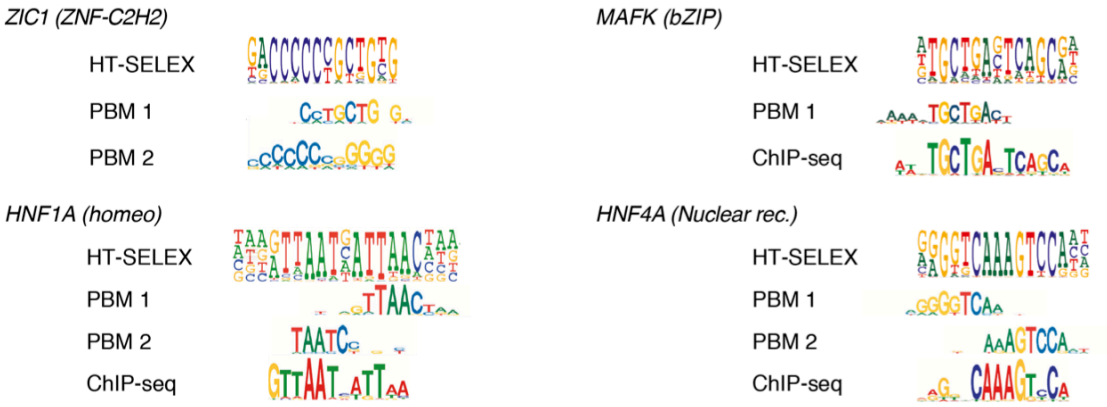


Figure 15 | Comparison of PBM, HT-SELEX and ChIP-seq data for TFs with long specificities
Even though the PBM method is capable of generation of accurate PWM models for certain TFs, the HT-SELEX outperforms PBM in the cases where the TFs recognize long target sites. In the four indicated cases, the HT-SELEX approach leads to the generation of long target sites (>10 bp), which are similar to the sites generated from analysis of ChIP-seq data in the three cases where such data was available. The PBM method on the other hand leads to the generation of partial binding specificity models, because the method is limited by the coverage of long target sites. The figure is adapted from[72].

Additional drawbacks include the fact that PBM method operates a single sample at a time, so the throughput is low, it is fairly expensive, and it has a relatively high (µg level) demand for purified protein. The data generated with HT-SELEX and PBM methods have been compared against each other in Study II and in two independent studies[171,221]. Based on these findings, the results are mostly similar, but on closer analysis the HT-SELEX performs better in the prediction of *in vivo* targets and for TFs that have long target sites. On the other hand, PBMs give more robust measures for modeling of TF-specificities as 8-mers, which is likely due to too low sequencing coverage of the HT-SELEX ligands[171,221].

The main benefit that HT-SELEX has over B1H is that in B1H the number of molecules that undergo selection is practically limited by bacterial transformation efficiency (~10,000,000 molecules). The HT-SELEX on the other hand can easily use libraries of 100 ng of DNA, which has several thousand times the number of the molecules used in a B1H experiment. In practice, the difference is even larger as the B1H clones are counter-selected before the actual selection step[221].

Traditional SELEX was severely limited by the sequencing output that could be gained from Sanger sequencing. This was remedied to some extent by using a protocol that concatenated ligands from the selected pool into continuous fragments that could then be cloned and sequenced to gain higher coverage of ligands with fewer sequencing reads[222] and finally by adaption of the SELEX to massively parallel sequencing simultaneously by us and others (Bind-n-seq[223], SELEX-seq[224] and HT-SELEX[225]).

6.1.1.2 Cooperative binding of TFs

All three studies showed instances where the TFs bind DNA cooperatively as homomultimeric complexes. Cooperative binding of TFs also often leads to recognition of composite sites where the two binding motifs overlap. The formation of composite site recognizing complexes is in most cases likely to be mediated through the DNA shape, where the binding of one of the TFs changes the shape of the DNA in a way that promotes the binding of the other TF to the adjacent position. This DNA-mediated conformational compatibility was observed commonly between many homodimeric pairs and is similar to previous observations based on analyses of crystal structures of DNA binding multiprotein complexes, such as the interferon-beta enhanceosome[125,226]. Additionally, experimental analyses performed on POU2F1 have shown that the two DBDs of this TF can cooperate through DNA even if the peptide-link between them is severed[161]. In many cases we also observed longer-range cooperativity (Figure 16b) that is likely to be based on DNA shape although more indirectly as it is mediated through oscillation of the vibrations in the DNA[160].

Even though many of the homodimers captured by the HT-SELEX, are mediated through a combination of protein-protein interactions with either of the DNA mediated mechanisms, for example the previously characterized homodimeric binding site of ETS1, the structure of which has been solved using X-ray crystallography[103,104], the results suggested that the cooperative interactions of the DNA binding TFs are dominated by the DNA-shape based mechanisms.

Orientation and spacing preferences mediated through DNA shape based cooperativity appear to change more often during the evolution than the primary specificities of the TFs. Based on the findings in Studies II and III, many of the paralogous TFs that bind highly similar individual sites formed homodimeric TF-complexes with different orientation and spacing preferences. This kind of divergence has taken place multiple times during the evolution and in many different structural classes, such as in ETS-, FOX- and TBX families. Experiments in all of the described studies analysed TFs only in individual context and the results were limited to homodimers, however there is no reason why the DNA-shape based cooperative behaviour should be limited to homodimers and thus the mechanism is very likely to be used commonly also in heterodimeric contexts.

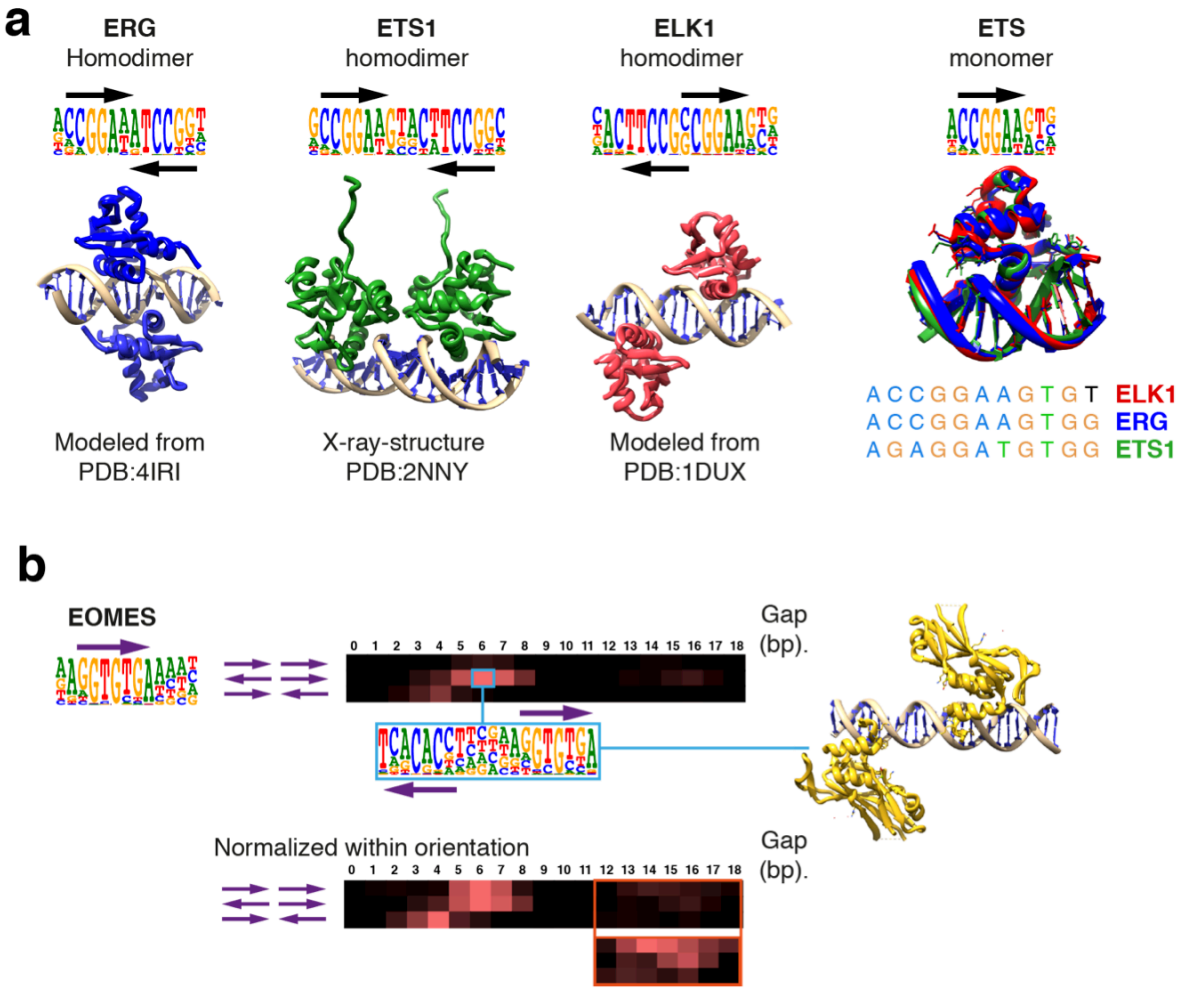


Figure 16 | DNA shape based cooperativity (DNA allostery)

a) Divergence of multimeric specificity in ETS class I TFs; While all of the class I ETS TFs bind highly similar specificity when binding the DNA individually (example PWM on top right corner), the ETS TFs ERG, ETS1 and ELK1 have evolved to recognize different homodimeric orientation and/or spacing preferences (black arrows). In the ETS1 structure contacts are formed between the DBDs, while in the case of ERG and ELK1 the cooperativity is likely to be mediated entirely through DNA shape. On the right side of the figure the three structures of ETS DBDs have been aligned based on the nucleotides CCGGAA or matching positions in the ETS1. Structural alignment shows that the DNA has different shape in all of the structures. Note however that because the ETS1 has been crystallized with a different (low-affinity) target-site, it is hard to know how much of the shape difference is due to differences between the proteins or the inherent DNA shape. Structure of ETS1 homodimer has been solved using X-ray crystallography while the ERG and ELK1 structures are schematic models based on fitting the monomeric crystal structures of the two TFs into a idealized structural model of B-DNA[103,227,228]. **b)** Another kind of DNA allostery occurs commonly when two TFs are located longer distance away from each other. *Above* T-box TF EOMES (monomer PWM on *left*, the site is also indicated as s violet arrow in the heatmaps) displays preferential binding of the two TFs to certain spacing and orientation combinations that are shown as heatmap visualizations where the red color indicates the fraction of the highest position. In the most preferred combination of spacing and orientation the sites have 6 bases long gap between the individual binding sites that are oriented tail to tail. This site is shown both as a PWM and as a schematic model based on crystal structure of paralogous TF TBX5 (PDB:2X6V)[229]. *Below,* normalizing the read counts within each orientation shows that EOMES sites display spacing preferences in all of the orientations, and that at least in two of them this phenomenon displays periodic behavior, as there is also another preferentially bound region that is located approximately ten bases further from the first one. The red-bordered inset shows all spacings and orientations occurring in this region that are normalized against the maximum value (as in the upper heatmap but normalized to preferences occurring in this region).

The shape based mechanisms offer an easy way to evolutionary divergence of paralogous TFs. By using this mechanism, the TFs can retain their individual binding specificities and instead diverge by adjusting the way that the TF-DNA interaction changes the shape of the DNA. The shape change then causes the paralogs to recognize partially separate subsets of target sites that either of the TFs recognizes when binding with homo- or heterodimeric binding partners.

6.1.1.3 Evolutionary perspective

Study III that generated new data for the binding specificities of the fruit fly TFs and compared the data to data of Study II found surprisingly high conservation of TF DNA-binding specificities. A possible explanation for this is that the functions of a TF are tied to a multitude of its target sites, and thus evolution of new specificities is constrained not just by the demands of the protein fold. On the other hand, a mutant TF that recognizes a novel specificity is in most of the circumstances likely to act like a loose cannon, recognizing new target sites that are located stochastically in random locations of the genome.

All in all, the combined data from all of the three studies reinforce the notion that even though there are many known instances of proteins that have evolved into new roles in specific animals, evolution of complex multicellular organisms happens to a large extent at the level of regulatory elements and not in the protein coding sequences. This makes a lot of sense; proteins are intricate and fascinating things that have been optimized for their roles over time spans of hundreds of millions of years, but in the end each of them performs usually just one or few tasks.

Beneficial mutation on a protein-coding gene itself is very unlikely because of the general principles of protein structure; any substitution of an amino acid to another is most likely to be either functionally silent or harmful. Additionally, it is very unlikely that a change in any given individual component would manifest as a fitness increase, as it is just one out of thousands and the entire organism is composed of millions of cells belonging to hundreds of different cell types. Mutations in regulatory elements on the other hand do not change the components of the cells but their quantities. In most cases, such a change would take place in enhancers and due to this the expression level change would be limited to a specific range of cell types and leave the others completely unaffected. Therefore, the probability that the organism tolerates the mutation is much higher when it occurs in regulatory elements rather than protein coding regions of the genes.

7 Conclusions, remarks and future prospects

The studies described in this thesis represent an important advance in the field of transcriptional gene regulation. An understanding of gene regulation is of vital importance to essentially all fields of molecular biology and medicine, and the effect of these studies reaches even beyond. Furthermore, as the studies also showed the astonishing level of conservation of TF sequence specificities between the fruit fly and human, the results are not limited to our own- or even closely related species such as other mammals, but apply well to all forms of animals.

It is impossible to understand gene regulation of animals without information about their gene regulatory components and thus the major findings presented in these studies are not any singular things, but the entire generated datasets composed of hundreds of models for TF specificities.

7.1.1.1 Future prospects

More studies have been planned or are already under preparation. Besides the natural follow-up through increasing the coverage of the TF model collections there are other important aspects that need and can be addressed with approaches that are similar to the HT-SELEX, for example: Effect of CpG methylation to DNA binding; Formation and specificities of heterodimeric TF complexes and; studying the sequence specificity of RNA binding proteins. As seen throughout this thesis, structural information about the TFs and their DBDs is indispensable for understanding their function. Because structures of many kinds of DBDs are still unknown, there is a demand for large number of X-ray crystallography experiments to determine the structures of DBDs binding into their high affinity target site.

8 ACKNOWLEDGEMENTS

The whole thesis work was conducted in the laboratory of Prof. Jussi Taipale, first in University of Helsinki and then in Karolinska Institutet.

First of all I'd like to express my thanks to my supervisors; Professor Jussi Taipale who is a good, smart and fair boss and Dr. Minna Taipale, who is usually the general voice of reason and sanity.

The second level of thanks is reserved for the following colleagues who shared the frontline as siblings in arms when we were fighting with one or more of the projects or manuscripts; Drs. Jian Yan, Kazuhiro Nitta, Yimeng Yin, Estefania Mondragon, Gonghong Wei, Martin Enge, Teemu Kivioja, Thomas Whitington and Ekaterina Morgunova and all of the other co-authors.

Special thanks going to the persons who helped helped by reading this thesis and giving feedback for it: Drs. Minna Taipale, Bernhard Schmierer, Lennart Nilsson and Grzegorz Raszewski, Msc. Anni Jolma and Mr. Mikko Koivu-Jolma.

Rest of the current and past members of the Taipale lab from both KI and University of Helsinki, there are just too many of you to get through within the time I had remaining while doing this part.

My previous boss Esa Kuismanen, who sold me to Jussi for a price of one "Suffeli" chocolate-bar.

My family, their spouses and F1 generation.

Rest of my friends and otherwise important mammalian individuals in (pseudo) random order (Ok. I admit iterating the sort a couple of times until I got someone very important to the first spot): Kummimummi, Tanja, Valtteri, Aleksi, Veikko, Olof, Anna(x3), Marko, Jere, Marcel, Tumppi, Mikki, Rudolf, Anna, Greg, Saana*, Nenni, Billy, Dina, Thatiane, Minna, Milla, Matti, Jelena, Johan, Flammie, Ola, Markus, Valeria, Mia, Terttu, Tarmo*, Lauri and those creatures who were not mentioned separately either because of space or memory limitations such as rest of the karate group members or the great apes from HYRMY (University of Helsinki academic metalheads organization).

**Canis lupus familiaris*

9 REFERENCES

1 Grummt, I. Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. *Genes Dev* **17**, 1691-1702, doi:10.1101/gad.1098503R (2003).

2 Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M. & Pagano, A. The expanding RNA polymerase III transcriptome. *Trends in genetics : TIG* **23**, 614-622, doi:10.1016/j.tig.2007.09.001 (2007).

3 Alberts, B. *et al. Molecular Biology of the Cell*. 4th edn, (Garland, 2002).

4 Chen, T. & Dent, S. Y. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nature reviews. Genetics* **15**, 93-106, doi:10.1038/nrg3607 (2014).

5 Rosenfeld, M. G., Lunyak, V. V. & Glass, C. K. Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes Dev* **20**, 1405-1428, doi:10.1101/gad.1424806 (2006).

6 Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**, 29-59, doi:10.1146/annurev.genom.7.080505.115623 (2006).

7 Beck, M. *et al.* The quantitative proteome of a human cell line. *Molecular systems biology* **7**, 549, doi:msb201182 10.1038/msb.2011.82 (2012).

8 Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 19428-19433, doi:10.1073/pnas.0709013104 (2007).

9 International Human Genome Sequencing, C. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

10 Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nature reviews. Genetics* **14**, 288-295, doi:10.1038/nrg3458 (2013).

11 Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 6131-6138, doi:10.1073/pnas.1318948111 (2014).

12 Hancock, R. The crowded nucleus. *Int Rev Cell Mol Biol* **307**, 15-26, doi:B978-0-12-800046-5.00002-3 10.1016/B978-0-12-800046-5.00002-3 (2014).

13 Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402-405, doi:10.1038/nature13986 (2014).

14 Rothbart, S. B. & Strahl, B. D. Interpreting the language of histone and DNA modifications. *Biochimica et biophysica acta* **1839**, 627-643, doi:S1874-9399(14)00052-2 10.1016/j.bbagrm.2014.03.001 (2014).

15 Swygert, S. G. & Peterson, C. L. Chromatin dynamics: interplay between remodeling enzymes and histone modifications. *Biochimica et biophysica acta* **1839**, 728-736, doi:S1874-9399(14)00034-0 10.1016/j.bbagrm.2014.02.013 (2014).

16 Luger, K., Dechassa, M. L. & Tremethick, D. J. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature reviews. Molecular cell biology* **13**, 436-447, doi:10.1038/nrm3382 (2012).

17 Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nature reviews. Molecular cell biology* **7**, 540-546, doi:10.1038/nrm1938 (2006).

18 Ziller, M. J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS genetics* **7**, e1002389, doi:10.1371/journal.pgen.1002389 (2011).

19 Roloff, T. C., Ropers, H. H. & Nuber, U. A. Comparative study of methyl-CpG-binding domain proteins. *BMC genomics* **4**, 1 (2003).

20 Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V. & Jordan, I. K. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3**, 462-474 (2012).

21 Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature reviews. Molecular cell biology* **14**, 341-356, doi:10.1038/nrm3589 (2013).

22 Rowe, H. M. *et al.* De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. *Development* **140**, 519-529, doi:10.1242/dev.087585 (2013).

23 Gregg, C. *et al.* High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329**, 643-648, doi:science.1190830 10.1126/science.1190830 (2010).

24 Comb, M. & Goodman, H. M. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. *Nucleic acids research* **18**, 3975-3982, doi:10.1093/nar/18.13.3975 (1990).

25 Iguchi-Ariga, S. M. & Schaffner, W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Genes Dev* **3**, 612-619 (1989).

26 Prendergast, G. C. & Ziff, E. B. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science* **251**, 186-189 (1991).

27 Mann, I. K. *et al.* CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. *Genome research* **23**, 988-997, doi:10.1101/gr.146654.112 (2013).

28 Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nature reviews. Genetics* **10**, 161-172, doi:10.1038/nrg2522 (2009).

29 Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell research* **21**, 381-395, doi:cr201122 10.1038/cr.2011.22 (2011).

30 Burgess, R. J. & Zhang, Z. Histone chaperones in nucleosome assembly and human disease. *Nat Struct Mol Biol* **20**, 14-22, doi:10.1038/nsmb.2461 (2013).

31 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

32 Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *Journal of molecular biology* **319**, 1097-1113, doi:10.1016/s0022-2836(02)00386-8 (2002).

33 Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS genetics* **4**, e1000138, doi:10.1371/journal.pgen.1000138 (2008).

34 Reynolds, N., O'Shaughnessy, A. & Hendrich, B. Transcriptional repressors: multifaceted regulators of gene expression. *Development* **140**, 505-512, doi:10.1242/dev.083105 (2013).

35 Clapier, C. R. & Cairns, B. R. The biology of chromatin remodeling complexes. *Annu Rev Biochem* **78**, 273-304, doi:10.1146/annurev.biochem.77.062706.153223 (2009).

36 Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**, 267-273, doi:10.1038/nsmb.2506 (2013).

37 Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology* **276**, 19-42, doi:10.1006/jmbi.1997.1494 (1998).

38 Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**, 2227-2241, doi:10.1101/gad.176826.111 (2011).

39 Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412-420, doi:10.1038/364412a0 (1993).

40 Cirillo, L. A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Molecular cell* **9**, 279-289 (2002).

41 Jerabek, S., Merino, F., Scholer, H. R. & Cojocaru, V. OCT4: dynamic DNA binding pioneers stem cell pluripotency. *Biochimica et biophysica acta* **1839**, 138-154, doi:10.1016/j.bbagrm.2013.10.001 (2014).

42 Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:10.1016/j.cell.2006.02.043 (2006).

43 Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature biotechnology*, doi:10.1038/nbt.3128 (2015).

44 Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237-240, doi:10.1038/nature08674 (2010).

45 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

46 Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol* **3**, REVIEWS0004 (2002).

47 Roy, A. L. & Singer, D. S. Core promoters in transcription: old problem, new insights. *Trends in biochemical sciences* **40**, 165-171, doi:10.1016/j.tibs.2015.01.007 (2015).

48 Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and

modes of higher order chromatin organization. *Cell research* **22**, 490-503, doi:10.1038/cr.2012.15 (2012).

49 Jing, H. *et al.* Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Molecular cell* **29**, 232-242, doi:10.1016/j.molcel.2007.11.020 (2008).

50 Doyle, B., Fudenberg, G., Imakaev, M. & Mirny, L. A. Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS computational biology* **10**, e1003867, doi:10.1371/journal.pcbi.1003867 (2014).

51 He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E2191-2199, doi:10.1073/pnas.1320308111 (2014).

52 Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).

53 Malnic, B., Godfrey, P. A. & Buck, L. B. The human olfactory receptor gene family. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2584-2589 (2004).

54 Lomvardas, S. *et al.* Interchromosomal interactions and olfactory receptor choice. *Cell* **126**, 403-413, doi:10.1016/j.cell.2006.06.035 (2006).

55 Marsman, J. & Horsfield, J. A. Long distance relationships: enhancer-promoter communication and dynamic gene transcription. *Biochimica et biophysica acta* **1819**, 1217-1227, doi:10.1016/j.bbagrm.2012.10.008 (2012).

56 Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387-396 (1999).

57 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:nature11082 10.1038/nature11082 (2012).

58 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385, doi:nature11049 10.1038/nature11049 (2012).

59 Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680, doi:S0092-8674(14)01497-4 10.1016/j.cell.2014.11.021 (2014).

60 Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics* **15**, 234-246, doi:10.1038/nrg3663 (2014).

61 Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome research* **22**, 1680-1688, doi:10.1101/gr.136101.111 (2012).

62 Kornberg, R. D. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 12955-12961, doi:0704138104 10.1073/pnas.0704138104 (2007).

63 Plaschka, C. *et al.* Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature* **518**, 376-380, doi:10.1038/nature14229 (2015).

64 Poss, Z. C., Ebmeier, C. C. & Taatjes, D. J. The Mediator complex and transcription regulation. *Crit Rev Biochem Mol Biol* **48**, 575-608, doi:10.3109/10409238.2013.840259 (2013).

65 Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801-813, doi:10.1016/j.cell.2013.07.034 (2013).

66 Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 996-1001, doi:10.1073/pnas.1317788111 (2014).

67 Chalkley, G. E. & Verrijzer, C. P. DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. *EMBO J* **18**, 4835-4845, doi:10.1093/emboj/18.17.4835 (1999).

68 Wassarman, D. A. & Sauer, F. TAF(II)250: a transcription toolbox. *Journal of cell science* **114**, 2895-2902 (2001).

69 Vermeulen, M. *et al.* Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58-69, doi:10.1016/j.cell.2007.08.016 (2007).

70 Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes Dev* **10**, 2657-2683 (1996).

71 Kwak, H. & Lis, J. T. Control of transcriptional elongation. *Annual review of genetics* **47**, 483-508, doi:10.1146/annurev-genet-110711-155440 (2013).

72 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339, doi:S0092-8674(12)01496-1 10.1016/j.cell.2012.12.009 (2013).

73 Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723, doi:1162327 10.1126/science.1162327 (2009).

74 Chan, C. P., Kok, K. H. & Jin, D. Y. CREB3 subfamily transcription factors are not created equal: Recent insights from global analyses and animal models. *Cell Biosci* **1**, 6, doi:2045-3701-1-6 10.1186/2045-3701-1-6 (2011).

75 Aaronson, D. S. & Horvath, C. M. A road map for those who don't know JAK-STAT. *Science* **296**, 1653-1655, doi:10.1126/science.1071545 (2002).

76 Dorsey, M. J. *et al.* B-ATF: a novel human bZIP protein that associates with members of the AP-1 transcription factor family. *Oncogene* **11**, 2255-2265 (1995).

77 Frietze, S. & Farnham, P. J. Transcription factor effector domains. *Sub-cellular biochemistry* **52**, 261-277, doi:10.1007/978-90-481-9069-0_12 (2011).

78 Scully, K. M. *et al.* Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science* **290**, 1127-1131 (2000).

79 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-263, doi:nrg2538 10.1038/nrg2538 (2009).

80 Fulton, D. L. *et al.* TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol* **10**, R29, doi:10.1186/gb-2009-10-3-r29 (2009).

81 Wei, G. H. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J*, doi:emboj2010106 10.1038/emboj.2010.106 (2010).

82 Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266-1276 (2008).

83 Kraus, P. & Lufkin, T. Dlx homeobox gene control of mammalian limb and craniofacial development. *American journal of medical genetics. Part A* **140**, 1366-1374, doi:10.1002/ajmg.a.31252 (2006).

84 Burglin, T. R. Homeodomain subtypes and functional diversity. *Sub-cellular biochemistry* **52**, 95-122, doi:10.1007/978-90-481-9069-0_5 (2011).

85 Brayer, K. J. & Segal, D. J. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem Biophys* **50**, 111-131, doi:10.1007/s12013-008-9008-5 (2008).

86 Laity, J. H., Lee, B. M. & Wright, P. E. Zinc finger proteins: new insights into structural and functional diversity. *Current opinion in structural biology* **11**, 39-46 (2001).

87 Tadepally, H. D., Burger, G. & Aubry, M. Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. *BMC evolutionary biology* **8**, 176, doi:10.1186/1471-2148-8-176 (2008).

88 Mukherjee, K. & Burglin, T. R. Comprehensive analysis of animal TALE homeobox genes: new conserved motifs and cases of accelerated evolution. *J Mol Evol* **65**, 137-153, doi:10.1007/s00239-006-0023-0 (2007).

89 Anderson, D. M. *et al.* Characterization of the DNA-binding properties of the Mohawk homeobox transcription factor. *J Biol Chem* **287**, 35351-35359, doi:10.1074/jbc.M112.399386 (2012).

90 Williams, T. M., Williams, M. E. & Innis, J. W. Range of HOX/TALE superclass associations and protein domain requirements for HOXA13:MEIS interaction. *Developmental biology* **277**, 457-471, doi:S0012-1606(04)00719-5 10.1016/j.ydbio.2004.10.004 (2005).

91 Shen, W. F. *et al.* AbdB-like Hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. *Mol Cell Biol* **17**, 6448-6458 (1997).

92 Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* **397**, 714-719, doi:10.1038/17833 (1999).

93 Jones, S. An overview of the basic helix-loop-helix proteins. *Genome Biol* **5**, 226, doi:10.1186/gb-2004-5-6-226 (2004).

94 Yoshida, T., Ohkumo, T., Ishibashi, S. & Yasuda, K. The 5'-AT-rich half-site of Maf recognition element: a functional target for bZIP transcription factor Maf. *Nucleic acids research* **33**, 3465-3478, doi:33/11/3465 10.1093/nar/gki653 (2005).

95 Deppmann, C. D., Alvania, R. S. & Taparowsky, E. J. Cross-species annotation of basic leucine zipper factor interactions: Insight into the evolution of closed interaction networks. *Molecular biology and evolution* **23**, 1480-1492, doi:msl022 10.1093/molbev/msl022 (2006).

96 Newman, J. R. & Keating, A. E. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* **300**, 2097-2101, doi:10.1126/science.1084648 (2003).

97 Tuteja, G. & Kaestner, K. H. Forkhead transcription factors II. *Cell* **131**, 192, doi:10.1016/j.cell.2007.09.016 (2007).

98 Brent, M. M., Anand, R. & Marmorstein, R. Structural basis for DNA recognition by FoxO1 and its regulation by posttranslational modification. *Structure* **16**, 1407-1416, doi:S0969-2126(08)00291-8 10.1016/j.str.2008.06.013 (2008).

99 Schlake, T., Schorpp, M., Nehls, M. & Boehm, T. The nude gene encodes a sequence-specific DNA binding protein with homologs in organisms that lack an anticipatory immune system. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 3842-3847 (1997).

100 Nakagawa, S., Gisselbrecht, S. S., Rogers, J. M., Hartl, D. L. & Bulyk, M. L. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 12349-12354, doi:10.1073/pnas.1310430110 (2013).

101 Aranda, A. & Pascual, A. Nuclear Hormone Receptors and Gene Expression. *Physiological Reviews* **81**, 1269-1304 (2001).

102 Seth, A. & Watson, D. K. ETS transcription factors and their emerging roles in human cancer. *European journal of cancer* **41**, 2462-2478, doi:10.1016/j.ejca.2005.08.013 (2005).

103 Babayeva, N. D. *et al.* Structural basis of Ets1 cooperative binding to palindromic sequences on stromelysin-1 promoter DNA. *Cell cycle* **9**, 3054-3062 (2010).

104 Lamber, E. P. *et al.* Regulation of the transcription factor Ets-1 by DNA-mediated homo-dimerization. *EMBO J* **27**, 2006-2017, doi:emboj2008117 10.1038/emboj.2008.117 (2008).

105 De Val, S. *et al.* Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell* **135**, 1053-1064, doi:S0092-8674(08)01387-1 10.1016/j.cell.2008.10.049 (2008).

106 Shiina, M. *et al.* A Novel Allosteric Mechanism on Protein-DNA Interactions underlying the Phosphorylation-Dependent Regulation of Ets1 Target Gene Expressions. *Journal of molecular biology*, doi:10.1016/j.jmb.2014.07.020 (2014).

107 Fitzsimmons, D. *et al.* Pax-5 (BSAP) recruits Ets proto-oncogene family proteins to form functional ternary complexes on a B-cell-specific promoter. *Genes Dev* **10**, 2198-2211 (1996).

108 Wegner, M. From head to toes: the multiple facets of Sox proteins. *Nucleic acids research* **27**, 1409-1420 (1999).

109 Naiche, L. A., Harrelson, Z., Kelly, R. G. & Papaioannou, V. E. T-box genes in vertebrate development. *Annual review of genetics* **39**, 219-239, doi:10.1146/annurev.genet.39.073003.105925 (2005).

110 Coll, M., Seidman, J. G. & Muller, C. W. Structure of the DNA-bound T-box domain of human TBX3, a transcription factor responsible for ulnar-mammary syndrome. *Structure* **10**, 343-356 (2002).

111 Conlon, F. L., Fairclough, L., Price, B. M., Casey, E. S. & Smith, J. C. Determinants of T box protein specificity. *Development* **128**, 3749-3758 (2001).

112 Ryan, A. K. & Rosenfeld, M. G. POU domain family values: flexibility, partnerships, and developmental codes. *Genes Dev* **11**, 1207-1225 (1997).

113 Esch, D. *et al.* A unique Oct4 interface is crucial for reprogramming to pluripotency. *Nat Cell Biol* **15**, 295-301, doi:http://www.nature.com/ncb/journal/v15/n3/abs/ncb2680.html - supplementary-information (2013).

114 Klemm, J. D., Rould, M. A., Aurora, R., Herr, W. & Pabo, C. O. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **77**, 21-32 (1994).

115 Ng, C. K. *et al.* Deciphering the Sox-Oct partner code by quantitative cooperativity measurements. *Nucleic acids research* **40**, 4933-4941, doi:10.1093/nar/gks153 (2012).

116 Trainor, C. D., Ghirlando, R. & Simpson, M. A. GATA zinc finger interactions modulate DNA binding and transactivation. *J Biol Chem* **275**, 28157-28166, doi:10.1074/jbc.M000020200 (2000).

117 Di Stefano, L., Jensen, M. R. & Helin, K. E2F7, a novel E2F featuring DP-independent repression of a subset of E2F-regulated genes. *EMBO J* **22**, 6289-6298, doi:10.1093/emboj/cdg613 (2003).

118 Tao, Y., Kassatly, R. F., Cress, W. D. & Horowitz, J. M. Subunit composition determines E2F DNA-binding site specificity. *Mol Cell Biol* **17**, 6994-7007 (1997).

119 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

120 Rao, A., Luo, C. & Hogan, P. G. Transcription factors of the NFAT family: regulation and function. *Annual review of immunology* **15**, 707-747, doi:10.1146/annurev.immunol.15.1.707 (1997).

121 Blake, J. A. & Ziman, M. R. Pax genes: regulators of lineage specification and progenitor cell maintenance. *Development* **141**, 737-751, doi:10.1242/dev.091785 (2014).

122 Paun, A. & Pitha, P. M. The IRF family, revisited. *Biochimie* **89**, 744-753, doi:10.1016/j.biochi.2007.01.014 (2007).

123 Escalante, C. R., Nistal-Villan, E., Shen, L., Garcia-Sastre, A. & Aggarwal, A. K. Structure of IRF-3 bound to the PRDIII-I regulatory element of the human interferon-beta enhancer. *Molecular cell* **26**, 703-716, doi:10.1016/j.molcel.2007.04.022 (2007).

124 Fujii, Y. *et al.* Crystal structure of an IRF-DNA complex reveals novel DNA recognition and cooperative binding to a tandem repeat of core sequences. *EMBO J* **18**, 5028-5041, doi:10.1093/emboj/18.18.5028 (1999).

125 Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. *Cell* **129**, 1111-1123, doi:10.1016/j.cell.2007.05.019 (2007).

126 Massagué, J., Seoane, J. & Wotton, D. Smad transcription factors. *Genes & Development* **19**, 2783-2810, doi:10.1101/gad.1350705 (2005).

127 Chai, J. *et al.* Features of a Smad3 MH1-DNA complex. Roles of water and zinc in DNA binding. *J Biol Chem* **278**, 20327-20331, doi:10.1074/jbc.C300134200 (2003).

128 Iyaguchi, D., Yao, M., Watanabe, N., Nishihira, J. & Tanaka, I. DNA recognition mechanism of the ONECUT homeodomain of transcription factor HNF-6. *Structure* **15**, 75-83, doi:10.1016/j.str.2006.11.004 (2007).

129 Gajiwala, K. S. *et al.* Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature* **403**, 916-921, doi:10.1038/35002634 (2000).

130 Emery, P. *et al.* A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity. *Mol. Cell. Biol.* **16**, 4486-4494 (1996).

131 Bromberg, J. F. Activation of STAT proteins and growth control. *BioEssays : news and reviews in molecular, cellular and developmental biology* **23**, 161-169, doi:10.1002/1521-1878(200102)23:2<161::AID-BIES1023>3.0.CO;2-0 (2001).

132 Ehret, G. B. *et al.* DNA binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites. *J Biol Chem* **276**, 6675-6688, doi:10.1074/jbc.M001748200 (2001).

133 Chen, X. *et al.* Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. *Cell* **93**, 827-839 (1998).

134 Shore, P. & Sharrocks, A. D. The MADS-box family of transcription factors. *European journal of biochemistry / FEBS* **229**, 1-13 (1995).

135 Pellegrini, L., Tan, S. & Richmond, T. J. Structure of serum response factor core bound to DNA. *Nature* **376**, 490-498, doi:10.1038/376490a0 (1995).

136 De Braekeleer, E. *et al.* RUNX1 translocations and fusion genes in malignant hemopathies. *Future oncology (London, England)* **7**, 77-91, doi:10.2217/fon.10.158 (2011).

137 Ito, Y., Bae, S.-C. & Chuang, L. S. H. The RUNX family: developmental regulators in cancer. *Nat Rev Cancer* **15**, 81-95, doi:10.1038/nrc3877 (2015).

138 Bravo, J., Li, Z., Speck, N. A. & Warren, A. J. The leukemia-associated AML1 (Runx1)--CBF beta complex functions as a DNA-induced molecular clamp. *Nature structural biology* **8**, 371-378, doi:10.1038/86264 (2001).

139 Bieging, K. T., Mello, S. S. & Attardi, L. D. Unravelling mechanisms of p53-mediated tumour suppression. *Nat Rev Cancer* **14**, 359-370, doi:10.1038/nrc3711 (2014).

140 Levrero, M. *et al.* The p53/p63/p73 family of transcription factors: overlapping and distinct functions. *Journal of cell science* **113**, 1661-1670 (2000).

141 Brandt, T., Petrovich, M., Joerger, A. C. & Veprintsev, D. B. Conservation of DNA-binding specificity and oligomerisation properties within the p53 family. *BMC genomics* **10**, 628, doi:10.1186/1471-2164-10-628 (2009).

142 Akiyama, Y., Hosoya, T., Poole, A. M. & Hotta, Y. The gcm-motif: a novel DNA-binding motif conserved in Drosophila and mammals. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 14912-14916 (1996).

143 Cohen, S. X. *et al.* Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. *The EMBO journal* **22**, 1835-1845, doi:10.1093/emboj/cdg182 (2003).

144 Slutsky, M. & Mirny, L. A. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophysical journal* **87**, 4021-4035, doi:10.1529/biophysj.104.050765 (2004).

145 Sekiya, T., Muthurajan, U. M., Luger, K., Tulin, A. V. & Zaret, K. S. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev* **23**, 804-809, doi:10.1101/gad.1775509 (2009).

146 von Hippel, P. H. & Berg, O. G. Facilitated target location in biological systems. *J Biol Chem* **264**, 675-678 (1989).

147 Elf, J., Li, G. W. & Xie, X. S. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316**, 1191-1194, doi:316/5828/1191 10.1126/science.1141967 (2007).

148 Hammar, P. *et al.* The lac repressor displays facilitated diffusion in living cells. *Science* **336**, 1595-1598, doi:336/6088/1595 10.1126/science.1221648 (2012).

149 Revyakin, A. *et al.* Transcription initiation by human RNA polymerase II visualized at single-molecule resolution. *Genes Dev* **26**, 1691-1702, doi:gad.194936.112 10.1101/gad.194936.112 (2012).

150 Chen, J. *et al.* Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell* **156**, 1274-1285, doi:S0092-8674(14)00197-4 10.1016/j.cell.2014.01.062 (2014).

151 Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* **79**, 233-269, doi:10.1146/annurev-biochem-060408-091030 (2010).

152 Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248-1253, doi:nature08473 10.1038/nature08473 (2009).

153 Sponer, J., Leszczynski, J. & Hobza, P. Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers* **61**, 3-31, doi:10.1002/1097-0282(2001)61:1<3::AID-BIP10048>3.0.CO;2-4 (2001).

154 Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 11163-11168 (1998).

155 Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences* **39**, 381-399, doi:S0968-0004(14)00121-2 10.1016/j.tibs.2014.07.002 (2014).

156 Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48, doi:10.1186/gb-2012-13-9-r48 (2012).

157 Wasson, T. & Hartemink, A. J. An ensemble model of competitive multi-factor binding of the genome. *Genome research* **19**, 2101-2112, doi:10.1101/gr.093450.109 (2009).

158 Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 22534-22539, doi:10.1073/pnas.0913805107 (2010).

159 Massari, M. E. & Murre, C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol Cell Biol* **20**, 429-440 (2000).

160 Kim, S. *et al.* Probing allostery through DNA. *Science* **339**, 816-819, doi:339/6121/816 10.1126/science.1229223 (2013).

161 Klemm, J. D. & Pabo, C. O. Oct-1 POU domain-DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes & development* **10**, 27-36 (1996).

162 Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of cellular biochemistry* **94**, 890-898, doi:10.1002/jcb.20352 (2005).

163 Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature reviews. Genetics* **15**, 453-468, doi:10.1038/nrg3684 (2014).

164 Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quantitative biology* **1**, 115-130, doi:10.1007/s40484-013-0012-4 (2013).

165 Weirauch, M. T. *et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology* **31**, 126-134, doi:10.1038/nbt.2486 (2013).

166 Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic acids research* **13**, 3021-3030 (1985).

167 Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic acids research* **10**, 2997-3011, doi:10.1093/nar/10.9.2997 (1982).

168 Benos, P. V., Bulyk, M. L. & Stormo, G. D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic acids research* **30**, 4442-4451 (2002).

169 Bulyk, M. L., Johnson, P. L. & Church, G. M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research* **30**, 1255-1261 (2002).

170 Mohibullah, N., Donner, A., Ippolito, J. A. & Williams, T. SELEX and missing phosphate contact analyses reveal flexibility within the AP-2[alpha] protein: DNA binding complex. *Nucleic acids research* **27**, 2760-2769, doi:gkc422 [pii] (1999).

171 Orenstein, Y. & Shamir, R. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic acids research* **42**, e63, doi:gku117 10.1093/nar/gku117 (2014).

172 Bi, Y., Kim, H., Gupta, R. & Davuluri, R. V. Tree-based position weight matrix approach to model transcription factor binding site profiles. *PloS one* **6**, e24210, doi:10.1371/journal.pone.0024210 (2011).

173 Roulet, E. *et al.* Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. *Journal of molecular biology* **297**, 833-848, doi:10.1006/jmbi.2000.3614 (2000).

174 Zhou, Q. & Liu, J. S. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **20**, 909-916, doi:10.1093/bioinformatics/bth006 (2004).

175 Mordelet, F., Horton, J., Hartemink, A. J., Engelhardt, B. E. & Gordan, R. Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics* **29**, i117-125, doi:10.1093/bioinformatics/btt221 (2013).

176 Sharon, E., Lubliner, S. & Segal, E. A feature-based approach to modeling protein-DNA interactions. *PLoS computational biology* **4**, e1000154, doi:10.1371/journal.pcbi.1000154 (2008).

177 Ellrott, K., Yang, C., Sladek, F. M. & Jiang, T. Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics* **18 Suppl 2**, S100-109 (2002).

178 Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome research* **16**, 123-131, doi:10.1101/gr.4074106 (2006).

179 Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* **17**, 877-885, doi:gr.5533506 10.1101/gr.5533506 (2007).

180 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218, doi: 10.1038/nmeth.2688 (2013).

181 Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-2309, doi:10.1126/science. 290/5500/2306 (2000).

182 Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419, doi:10.1016/j.cell.2011.11.013 (2011).

183 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799 295/5558/1306 (2002).

184 Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics* **38**, 1341-1347, doi:ng1891 10.1038/ng1891 (2006).

185 Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16**, 1299-1309, doi:gr.5571506 10.1101/gr.5571506 (2006).

186 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:326/5950/289 10.1126/science.1181369 (2009).

187 Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:nature08497 10.1038/nature08497 (2009).

188 Wong, M. L. & Medrano, J. F. Real-time PCR for mRNA quantitation. *BioTechniques* **39**, 75-85 (2005).

189 Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics* **14**, 618-630, doi:10.1038/nrg3542 (2013).

190 Naylor, L. H. Reporter gene technology: the future looks bright. *Biochemical pharmacology* **58**, 749-757 (1999).

191 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077, doi:science.1232542 10.1126/science.1232542 (2013).

192 Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic acids research* **5**, 3157-3170 (1978).

193 Hampshire, A. J., Rusling, D. A., Broughton-Head, V. J. & Fox, K. R. Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. *Methods* **42**, 128-140, doi:S1046-2023(07)00005-9 10.1016/j.ymeth.2007.01.002 (2007).

194 Jain, S. S. & Tullius, T. D. Footprinting protein-DNA complexes using the hydroxyl radical. *Nat Protoc* **3**, 1092-1100 (2008).

195 Oehler, S., Alex, R. & Barker, A. Is nitrocellulose filter binding really a universal assay for protein-DNA interactions? *Anal Biochem* **268**, 330-336, doi:S0003-2697(98)93056-1 10.1006/abio.1998.3056 (1999).

196 Fried, M. & Crothers, D. M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic acids research* **9**, 6505-6525 (1981).

197 Fried, M. G. & Liu, G. Molecular sequestration stabilizes CAP-DNA complexes during polyacrylamide gel electrophoresis. *Nucleic acids research* **22**, 5054-5059 (1994).

198 Garner, M. M. & Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic acids research* **9**, 3047-3060 (1981).

199 Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* **2**, 1849-1861, doi:nprot.2007.249 10.1038/nprot.2007.249 (2007).

200 Li, J. J. & Herskowitz, I. Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science* **262**, 1870-1874 (1993).

201 Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol Cell Biol* **9**, 2944-2949 (1989).

202 Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510 (1990).

203 Hallikas, O. & Taipale, J. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. *Nat Protoc* **1**, 215-222 (2006).

204 Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47-59 (2006).

205 Ladbury, J. E. Counting the calories to stay in the groove. *Structure* **3**, 635-639 (1995).

206 Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P. & Duhr, S. Molecular interaction studies using microscale thermophoresis. *Assay and drug development technologies* **9**, 342-353, doi:10.1089/adt.2011.0380 (2011).

207 Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233-237, doi:315/5809/233 10.1126/science.1131007 (2007).

208 Shumaker-Parry, J. S., Aebersold, R. & Campbell, C. T. Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy. *Anal Chem* **76**, 2071-2082, doi:10.1021/ac035159j (2004).

209 Berger, M. F. & Bulyk, M. L. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol* **338**, 245-260 (2006).

210 Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* **4**, 393-411, doi:nprot.2008.195 10.1038/nprot.2008.195 (2009).

211 Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**, 1429-1435 (2006).

212 Puckett, J. W. *et al.* Quantitative microarray profiling of DNA-binding molecules. *J Am Chem Soc* **129**, 12310-12319, doi:10.1021/ja0744899 (2007).

213 Meng, X., Brodsky, M. H. & Wolfe, S. A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature biotechnology* **23**, 988-994, doi:nbt1120 10.1038/nbt1120 (2005).

214 Noyes, M. B. *et al.* Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277-1289 (2008).

215 Noyes, M. B. *et al.* A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic acids research* **36**, 2547-2560, doi:gkn048 10.1093/nar/gkn048 (2008).

216 Enuameh, M. S. *et al.* Global analysis of Drosophila Cys(2)-His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome research* **23**, 928-940, doi:10.1101/gr.151472.112 (2013).

217 Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research* **36**, D102-106 (2008).

218 Vincentelli, R. *et al.* High-throughput protein expression screening and purification in Escherichia coli. *Methods* **55**, 65-72, doi:10.1016/j.ymeth.2011.08.010 (2011).

219 Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270-1282, doi:10.1016/j.cell.2011.10.053 (2011).

220 Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4**, doi:10.7554/eLife.04837 (2015).

221 Stormo, G. D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nature reviews. Genetics* **11**, 751-760, doi:10.1038/nrg2845 (2010).

222 Roulet, E. *et al.* High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nature biotechnology* **20**, 831-835 (2002).

223 Zykovich, A., Korf, I. & Segal, D. J. Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic acids research* **37**, e151, doi:gkp802 10.1093/nar/gkp802 (2009).

224 Zhao, Y., Granas, D. & Stormo, G. D. Inferring binding energies from selected binding sites. *PLoS computational biology* **5**, e1000590, doi:10.1371/journal.pcbi.1000590 (2009).

225 Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research* **20**, 861-873, doi:gr.100552.109 10.1101/gr.100552.109 (2010).

226 Panne, D., Maniatis, T. & Harrison, S. C. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. *EMBO J* **23**, 4384-4393, doi:7600453 10.1038/sj.emboj.7600453 (2004).

227 Cohen, S. X. *et al.* Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. *EMBO J* **22**, 1835-1845, doi:10.1093/emboj/cdg182 (2003).

228 Mo, Y., Vaessen, B., Johnston, K. & Marmorstein, R. Structure of the elk-1-DNA complex reveals how DNA-distal residues affect ETS domain recognition of DNA. *Nature structural biology* **7**, 292-297, doi:10.1038/74055 (2000).

229 Stirnimann, C. U., Ptchelkine, D., Grimm, C. & Muller, C. W. Structural basis of TBX5-DNA recognition: the T-box domain in its DNA-bound and -unbound form. *Journal of molecular biology* **400**, 71-81, doi:10.1016/j.jmb.2010.04.052 (2010).